В.В. Немцева (Плахотя), С.В.Чебанов

ЛИНГВОСТАТИСТИЧЕСКИЕ ПОСЛЕДСТВИЯ ОРФОГРАФИЧЕСКОЙ РЕФОРМЫ 1918 Г.

V.V. Nemtseva (Plakhotya), S.V.Chebanov

LINGUOSTATISTICAL CONSEQUENCES OF SPELLING REFORM OF 1918

Описывается суть реформы орфографии русского языка 1917-1918 гг. На основании изучения четырех массивов текстов (рассказы Чехова, роман Карлейля, правила грамматики Грота, стихи русских поэтом) изучается различие их лингвостатистических характеристик в дореформенной и пореформенной записи. Показывается небольшое, но достоверное, воспроизводимое и интерпретируемое различие этих записей по ряду параметров (число словоупотреблений, число словоформ, частота словоформ, энтропия и анэнтропия распределения). Ставится вопрос о необходимости учета этих различий при проведении лингвостатистических и стилеметрических исследований русской дореволюционной литературы. Рассматриваются общекультурные последствия выявленных различий.

In the article the essence of Russian spelling reform of the years 1917-1918 is described. The distinctions of linguostatistical characteristics in pre-reform and post-reform versions that are based on the four text file research (Chekhov's stories, the novel by Carlyle, grammar rules by Grooth, verses of Russian poets) are also demonstrated here. Small, but authentic, reproduced and interpreted distinctions of these versions by a number of parameters (number of word usage, number of word forms, frequency of word forms, entropy and anentropy of distributions) are shown. There is also raised the question of these distinctions consideration while carrying out linguostatistical and stylometric research of the Russian pre-revolutionary literature. General cultural consequences of the revealed distinctions are examined.

Ключевые слова: лингвостатистика, квантитативная лингвистика, орфография русского языка, реформа 1918 г., стилеметрия

Keywords: linguostatistics, quantitative linguistics, russian spelling, reform of 1918, stylometrics

---

Данная статья, как и предшествующие работы авторов (Плахотя, 2009, Плахотя, Чебанов, 2009) посвящена изучению влияния реформы орфографии 1917-1918 гг. на статистические характеристики русского письменного текста.

Реформа русской орфографии 1918 г.

Реформа заключалась в том, что (Декрет…, 1918):

- из алфавита исключались буквы **ѣ** (ять), **ѳ** (фита), **і** («и десятеричное»); вместо них должны употребляться, соответственно, **е, ф, и**;
- исключался твёрдый знак (**ъ**) на конце слов и частей сложных слов, но сохранялся в качестве разделительного знака (подъём, адъютант);
- изменялось правило написания приставок на «з»: теперь они кончались на «с» перед глухой и на «з» перед звонкой согласной и перед гласной (разбить, разораться, разступиться → разбить, разораться, но расступиться);
- в родительном и винительном падежах прилагательных и причастий окончания -аго, -яго заменялось на -ого, -его (напр. новаго → нового, лучшаго → лучшего, ранняго → раннего), в именительном и винительном падежах множественного числа женского и среднего родов -ыя, -ія — на -ые, -ие (если женского и среднего рода, то на «я», если мужского, то на «е»);
- словоформы женского рода множественного числа «онѣ», «однѣ», «однѣхъ», «однѣмъ», «однѣми» заменялись на «они», «одни», «одних», «одним», «одними»; словоформа родительного падежа единственного числа «ея» («нея») — на «ее» («нее») или «её» («неё»). К тому же, отмена «ѣ» создала необходимость в использовании буквы «ё», так как до реформы слова с «ѣ» произносились как «е», а с «е» как «ё». Поэтому без «ѣ» трудно было различать «все» и «всё», «вёдро» и «ведро», «всем» и «обо всём» и прочее.

Реформа ничего не говорила о судьбе редкой и выходящей из практического употребления ещё до 1917 года буквы **ѵ** (ижицы); на практике после реформы она также окончательно исчезла из алфавита.

Кроме того, реформа изменяла правила слитного и раздельного написания ряда выражений.

Новая орфография увеличила количество омографов в виду

- Исчезновения различий между одинаково звучащими словами, такими, как «**е**сть»/«**ѣ**сть», «м**и**ромъ»/«м**і**ромъ»/«м**ѵ**ромъ» и пр.

- Совпадения в современном русском языке форм существительных разных падежей. Например, в слове «поле» пореформенной орфографии совпали на письме именительный, винительный и предложный падежи, хорошо различавшиеся в Гротовской орфографии: «поле» (именительный, винительный) – въ полѣ (предложный). Устранение буквы «ѣ» (ять) из русского письма привело к неразличению указанных падежей в некоторых текстах.

Такая ситуация оказывается чревата как курьезными недоразумениями, так и серьезными семантическими сдвигами, которые явились основанием горячих дискуссий сторонников и противников реформы как до и во время ее проведения, так и впоследствии (Бодуэн-де-Куртенэ, 1912, Валентин, 1999, Досекин, 1917, Друговейко-Должанская, 1998, Иванов, 1905, Ильин, 1956, Кучеров, б.г., Лопатин, 2001, Мефодий, 1905, Мечковская., 1998, Реформа…, б.г., Сакулин, 1917, Сандомирская,1991, Сборник…, 2006, Троицкий, б.г., Тюрин, 2000, Чернышев,1912, Шагинян, 1917, Щерба, 1957, Языков, 1809).

Количественные характеристики последствий реформы рассматривались только с точки зрения полиграфии (так, переход на новую орфографию сокращает текст примерно на 3,4% со страницы, причем в первую очередь за счет ликвидации «ъ» в конце слов – в дореволюционной России ежегодно около 8,5 млн. страниц печатного текста приходились только на «ъ» – Успенский, 1974, с.192-195). Лингвостатистические изменения самого текста при этом не оценивались.

Методические проблемы изучения лингвостатистических последствий орфографической реформы 1918 г.

Развитие лингвостатистики и стилеметрии, появление компьютеров и совершенствование программных средств к ним дают надежду на относительно простой путь сопоставления количественных характеристик текста в до- и пореформенной орфографии.

Однако, на деле задача оказывается заметно более сложной как из-за технических проблем, связанных с оцифровкой дореформенного текста и его статистической обработкой, поскольку доступные пакеты лингвостатистических программ (такие, как LINDA) с дореформенной орфографией не работают, так и из-за проблемы выбора адекватной статистической модели текста.

Для проведения исследования осуществлялся отбор одних и тех же текстов в машиночитаемой форме в старой, и в новой орфографии. Основным материалом исследования послужили следующие массивы текстов в старой орфографии:

- Рассказы А.П.Чехова (24000 словоупотреблений – «В Москвѣ на Трубной площади», «Лошадиная фамилія», «Сильныя ощущенія», «Шуточка», «Perpetuum mobile», «В сараѣ», «На страстной недѣлѣ», «Нахлѣбники», «Унтеръ Пришибеевъ» т.3, с. 3-7, 13-20, 56-60, 65-77, 154-159, 172-177, 195-199, «Безпокойный гость», «Красавицы», «Мечты», «Слѣдователь», «Событіе», «Старый домъ» т.4, с. 3-22, 95-104, 164-170, 207-212 – Чехов А.П. ПСС в 23 т., Пг., Ком. нар. прос., 1918);

- Перевод на русский язык романа Т.Карлейля «Sartor Resartus. Жизнь и мысли Герр Тейфельсдрека» (73000 словоупотреблений – http://www.philol.msu.ru/~forlit/Pages/Word_versions/Carlyle/);

- Правила орфографии Я.К.Грота, 1885 (18000 словоупотреблений – Я.К.Грот «Правила русской орфографии и пунктуации» – http://www.geocities.com/kstati/texts/grot.html, сейчас перемещено на сайт http://members.shaw.ca/abaka/spell/texts/grot.htm);

- Массив, состоящий из стихотворений русских поэтов (23000 словоупотреблений – И.С.Аксаковъ: «Послѣ 1848 года», «Пусть сгибнетъ все», «Бываетъ такъ, что зодчій много лѣтъ», «Шоссе»; К.С.Аксаковъ: «Поэту Укорителю», «Семилѣтнему Л-у», «Петру», «Гуманисту», «Лже-духъ», «Литераторы натуралисты»; С.Т.Аксаковъ: «Посланіе къ А.Н.Майкову», «При вѣсти о грядущемъ освобожденіи крестьянъ»; П.А.Вяземскій: «Сельская церковь», «Замѣтка», «Меня „за ненависть къ журналамъ" судятъ строго», «Замѣтки», «Зачѣмъ глупцовъ ты задѣваешь?», «Намъ кое-что еще темно въ натурѣ», «Идутъ ли въ прокъ дѣла, иль плохо», «Попробуй съ рьянымъ неофитомъ», «De mortuis aut nihil, aut bene», «По мнѣ – онъ просто скучный враль», «Царь Горохъ», «Иному жизнь — одна игрушка», «Друзьямъ», «Старость», «Мужество», «Смиреніе», «Не тѣмъ трудна житейская задача», «Хорошіе люди», «Своей повадкой неусыпной», «Современныя замѣтки», «Вотъ мчится тройка удалая», «Одесса», «Александрійскій стихъ», «Литературная исповѣдь», «Plutôt être que paraître», «Одно сокровище», «Поминки», «Не рѣдко намъ — ктожъ не слыхалъ?—пеняли...», «Герои, славные мужья...»; Н.И.Гнѣдичъ: «На французскія революціонныя знамена въ Казанском соборѣ», «День моего рожденія»; Д.В.Давыдовъ: «Современная пѣсня», «Гр. П.А.Строгонову», «На монументъ Пожарскаго», «На смерть N. N», «Логика пьянаго», «Поэтическая женщина», «Бородинское поле», «Зайцевскому, поэту-моряку», «Листокъ»; Г.Р.Державинъ: «Эпитафія самому себѣ», «Уповающему на свою силу», «Христосъ», «Богъ»; М.А.Дмитріевъ: «Погибель гордыхъ», «Слѣпому поэту»; В.А.Жуковскій: «Старая пѣсня на новый ладъ», «Русская слава», «Пѣснь на присягу

государя наслѣдника», «Многолѣтие», «Народныя пѣсни»; И.И. Козловъ: «Моя молитва»; И.А.Крыловъ: «Подражаніе псалму XVII», «Ода. Вольное подражаніе псалму: „Господи, да не яростію Твоею“ и пр.», «Подражаніе псалму 71», «Подражаніе псалму 14», «Подражаніе псалму 96», «Подражаніе псалму 93», «Подражаніе псалму 51», «Подражаніе псалму 87»; А.Н.Майковъ: «4-е апрѣля 1866 года», «У гроба Грознаго», «Упраздненный монастырь», «Памяти Державина», «Посланіе в лагерь», «Я.П.Полонскому», «Изъ Аполлодора Гностика», «Оконченъ трудъ—ужь онъ мнѣ трудъ постылый», «’Не отставай отъ вѣка’ — лозунгъ лживый», «Перечитывая Пушкина», «Мы выросли въ суровой школѣ», «Сказаніе о 1812 годѣ», «Западная Русь», «М.Н. Каткову», «Ѳ.И. Тютчеву», «Казакъ», «Въ Айя-Софіи», «Судъ предковъ»; В.И.Майковъ: «Подражаніе псалму 1», «Подражаніе псалму 12», «Подражаніе псалму 41», «Подражаніе псалму 71», «Подражаніе псалму 81», «Подражаніе псалму 111», «Подражаніе псалму 136», «Подражаніе псалму: Внегда единоборствовалъ Давидъ на Голіафа», «О страшномъ судѣ», «Исходъ, Второзаконія», «Даніилъ»; А.Ѳ.Мерзляковъ: «Невинность», «Надгробная пѣснь», «Къ протекшему 1796 году», «Велизарій», «Среди долины ровныя», «Хоръ дѣтей маленькой Наташѣ», «Арсеній Несмѣлов», «Нищіе духом»; В.П.Петровъ: «На взятье Очакова», «На заключеніе мира со Швеціею»; Я.П.Полонскій: «Бранятъ», «На пути», «Въ нашъ вѣкъ», «Оттого что онъ вѣрить въ людей пересталъ», «По торжищамъ влача тяжелый крестъ поэта»; В.М.Пуришкевичъ «Русской женщинѣ»; Г.В.Сокольскій: «Къ протекшей юности»; Ѳ.И.Тютчевъ «Два голоса», «Олеговъ щитъ», «Декабристамъ», «На взятіе Варшавы», «Весна»; А.А.Фетъ: «1 Марта 1881 года», «15 Мая 1883 года», «Королевѣ эллиновъ Ольгѣ Константиновне», «Великому князю Константину Константиновичу», «Псевдо-поэту», «Когда кичливый умъ, измученный борьбою...», В.С.Филимоновъ «Надгробная пѣснь Державину»; Д.И.Хвостовъ «О наводненіи Петрополя»; И.И.Хемницеръ: «Метафизикъ», «Чужая бѣда», «Статуя», «Буквы», «Собака и мухи», «Дуракъ и тѣнь», «Хулитель стихотворства», «Остякъ и проѣзжій», «Кошка», «Счастливый мужъ», «Богачъ и бѣднякъ», «Львиный указъ», «Заяцъ, обойденный при произвожденіи», «Дворовая собака», «Великанъ и карлики», «Волчье разсужденье», «Желаніе кощея», «Паукъ и мухи», «Черви», «Привязанная собака»; А.С.Хомяковъ: «Не говорите: “То былое, то старина, то грѣхъ отцовъ”», «Давидъ», «Благочестивому меценату», «Подвигъ есть и въ сраженьи...», «По прочтеніи псалма», «Широка, необозрима...», «”Мы родъ избранный”, говорили»; С.П.Шевыревъ: «Устрой свой міръ внутри себя, Россія»; А.С.Шишковъ: «Шерстобой Андрей», «Езопъ», «Лисица и оселъ» –

http://tzvelodubov.livejournal.com/25847.html, а в новой на следующих сайтах: www.ru.wikisource.org, www.az.lib.ru, www.litera.ru, www.vyazemsky.ouc.ru, www.100stix2.ru, www.me4ta.su/books-lyrical/, www.stroki.net, www.svitlo.net, www.uucyc.ru, www.russianclassics.ru, www.maykov.ouc, www.community.livejournal.com, www.svitlo.by.ru). Этот массив особенно интересен, так как реформа орфографии для поэзии особенно чувствительна. Для этого сначала были найдены в машиночитаемой форме стихи в старой орфографии, а потом искались такие же в новой.

- Массив, в котором все указанные тексты были объединены. Его общий объем составил около 140 тысяч словоупотреблений.

Программа «URS Unique Record Set Management utility» использовалась для составления частотного словаря словоформ каждого из указанных массивов. Для импорта в URS пришлось заменять все буквы старой азбуки, отсутствующие в современной, на русские буквосочетания (ни английские, ни цифры в середине слова указанной программой не обрабатывались). Для большей наглядности Ять, Фита, И Десятеричное и Ижица были заменены на невстречающиеся в русском языке сочетания ЕЪТЮЛ, ФУБЯСЪ, ИЛЪЕК, ИЪЖИНЪ соответственно. После получения частотного списка была произведена обратная замена на исконные буквы. Для контроля весь массив был проверен вручную на правильность произведенных замен. Ввиду недоступности пригодного для написания программы списка правил лемматизации дореформенных словоформ, лексемы не выделялись (в том числе, и в текстах в современной орфографии) и их статистика не изучалась.

Полученные после такой обработки линвостатистические характеристики дореформенного текста не могут быть напрямую сопоставлены с результатами, полученными на текстах в современной орфографии с помощью других методов обработки. Это обстоятельство вызывает огорчение в связи с тем, что по последним имеются значительный по объему словарный материал и детальные математические модели, например, лингвостатистические особенности текстов А.П.Чехова детально описаны в работах А.О.Гребенникова (Гребенников, 1998). В виду этого обстоятельства для работы в качестве эталонного материала были выбраны рассказы Чехова, которые исследовались с помощью принятой методики; производилось и изучение тех же текстов в современной орфографии.

Следует заметить, что компьютерный анализ при составлении словарей не решал проблем, связанных с омонимией и омографией, которые, однако, для текстов в старой орфографии менее выражены.

При использовании данных методов обработки текстов были определенные искажения в величинах, вызванные ошибкой файнридера, которые требовали дальнейшей ручной сверки. Однако данными ошибками можно пренебречь без особого ущерба для общих выводов ввиду их редкости (например, для массива по Чехову около 150 ошибок распознавания на 24000 словоупотреблений, т.е.0,63%).

В результате было сформировано 10 массивов текстов, количественные характеристики которых приведены в табл. 1.

Таблица 1. Объемы исследованных массивов

	Количество словоупотреблений		Количество знаков			
			Без пробелов		С пробелами	
	До реформы	После реформы	До	После	До	После
А.П.Чехов	23688	23965	133308	126327	156697	154218
Т.Карлейль	73649	73653	447590	426664	524706	503783
Я.К.Грот	18513	18511	117059	111675	135922	130538
Стихотворения	22997	23115	124655	116321	148080	139745
Сводный	138847	139244	822612	780987	965405	928284

Следующим шагом было получение частотного словаря словоформ тех же произведений в современной орфографии. Для этого все данные текстов до- и пореформенного издания были перенесены в программу Excel 2007. С помощью стандартных функций этой программы строились полные частотные словари словоформ, а также определялись характеристики распределений.

Выбор таких характеристик определяется следующими соображениями.

Начиная с работ Дж.К.Ципфа (Zipf, 1932), известно, что в тексте слова (лексемы, словоформы) встречаются в резко неравном количестве. Распределения частот слов и аналогичные распределения в других предметных областях (Парето – в экономике, Брэдфорда – в социологии, Виллиса – в биологии и т.д.) оказалось удобно представлять в ранговой форме. В этом случае самому высокочастотному классу приписывается ранг k=1, следующему по частоте классу ранг k=2 и т.д., а численность класса ранга k P_k можно представлять как функцию ранга $P_k=f(k)$, а именно $P_{(k)}=A/k$, где A – определенный параметр (Шрейдер, 1996), описывающую некоторую гиперболу (отсюда используемое далее название *H*-распределение).

В 1970-2000-ые гг. предметом дискуссии оказалась математическая природа этих распределений. Одни исследователи (Алексеев, 1988, Бычков, 1986, Herdan, 1964) использовали для их характеристики обычные статистические параметры и предполагали существование у этих распределений репрезентативных совокупностей, другие трактовали

их как распределения с неопределенными центральными моментами всех порядков (Арапов, 1988, Арапов и др, 1975, Налимов, 1978, Орлов, 1970, 1996, Шрейдер, Шаров, 1982) и доказывали отсутствие у них характеристических совокупностей (Шрейдер, Шаров, 1982), Б.И.Кудрин представлял их как распределения степеней простых сомножителей численности генеральной совокупности (Кудрин, 2004, 2007), Г.Я.Мартыненко рассматривает как смесь двух распределений – высоко- и низкочастотных компонентов (Мартыненко, 1978) и т.д.

Наиболее принципиальным был вопрос о том, идет ли речь о гауссоподобном распределении случайной величины или о неопределенном распределении, то есть такого, для которого не выполняется центральная предельная теорема. Распределения второго типа в работах М.В.Арапова, А.П.Левича, В.В.Налимова, Ю.К.Орлова, Ю.А.Шрейдера и их сотрудников (Арапов, Шрейдер, 1977, Левич, б.г., Орлов, 1970, Шрейдер, 1967) связываются с объектами, обладающими целостностью, описывая их компонентный состав. С точки зрения логики такие совокупности являются собирательными множествами (Чебанов, 1979), компоненты которых характеризуются качественным разнообразием. В работах школы Б.И.Кудрина такие совокупности стали квалифицироваться как ценозы (Кудрин, 2006).

Так или иначе, вопросы, касающиеся математики ценозов, продолжают быть предметом жарких дискуссий. Не менее дискуссионным является и вопрос об оценке различия таких распределений, критериях достоверности обнаруживаемых различий. Строго говоря, любое различие обсуждаемых распределений с точки зрения ценологии является значимым (Кудрин и др., 2008, с.38). Статистика гауссоподобных распределений считает значимыми только различия, превышающие некоторую критическую величину.

Поскольку вопрос о природе лингвостатистических распределений является дискуссионным, для сравнения распределений в до- и пореформенной орфографии были отобраны критерии, используемые разными авторами, которые стоят порою на резко различных позициях.

В итоге для характеристики совокупностей были отобраны следующие показатели:

- Общее число словоупотреблений;

- Число словоформ с однократным употреблением и их доля относительно количества словоупотреблений;

- Энтропия $H = -\sum p_i log_2 p_i$, где i – ранг, p_i – соответствующая этому рангу частота появления словоформы для i=1÷k, где k – количество классов;

- Анэнтропия $A = -\sum log_2 p_i / n$, где i – ранг, p_i – соответствующая этому рангу частота появления словоформы для i=1÷k, где k – количество классов, а n – общее число словоупотреблений (Петров, 2005, с.57);

- Частота первых двенадцати самых частых слов и их доля;

- Мера различия, равная отношению частот i-го ранга в новой и старой орфографии – $K_i = f_{iнов}/f_{iстар}$

- Количество знаков с пробелами и без;

- Количество словоформ V/ количество словоупотреблений N;

- Пойнтер-точка – такая точка перегиба, в которой гипербола дискретно-непрерывно существует до этой точки, вырождаясь в ней Ω(R)=1 в прямую так, что далее все словоформы единичны (Кудрин, 2004, с.80). Пойнтер точка может быть найдена так: R=U-$K_н$, где $K_н$ – количество особей ноевой касты – количество уникальных слов, а U – общее количество словоупотреблений;

- Индекс разнообразия I = lgV/lgN, где V – объем словаря, N – объем выборки (формула предложена Г.Я Мартыненко);

- Ранговое среднее $r_{ср} = \sum r \bullet n_r / \sum n_r$, где r – ранг словоформы, n_r – соответствующая этому рангу частота словоформы (Мартыненко, 1978);

- Разности между этими данными в до- и пореформенных текстах.

С помощью этих показателей были охарактеризованы все массивы в новой и в старой орфографии, а их различия были проинтерпретированы с точки зрения анализа последствий реформы орфографии 1918г.

Обнаруженные различия

А) Общие характеристики

В итоге осуществления указанных выше расчетов были получены результаты, представленные в приводимых таблицах. Во всех приведенных ниже таблицах: X_1 – показатель для текстов в старой орфографии, X_2 – показатель для текстов в новой орфографии, X_1-X_2 – разность показателей одних и тех же текстов в старой и новой орфографии.

Для произведений А.П.Чехова были получены следующие результаты (табл.2).

Таблица 2. Характеристики массива по произведениям А.П.Чехова в до- и пореформенной орфографии.

Характеристика	X_1	X_2	$X_1 - X_2$
Количество словоупотреблений (U)	23688	23965	-277
Количество словоформ (каст, S)	8010	7977	33
S/U	0, 338	0,332	0,006
Количество словоформ с n_i=1	5620	5542	78
Пойнтер-точка *R*	2391	2436	-45
Ранговое среднее	1561,1	1539,4	21,7
Индекс разнообразия	0,892	0,890	0,002

Аналогичные данные, полученные по роману Т.Карлейя, приведены в табл.3.

Таблица 3. Характеристики массива по роману Т.Карлейя в до- и пореформенной орфографии.

Характеристика	X_1	X_2	$X_1 - X_2$
Количество словоупотреблений (U)	73649	73653	-4
Количество словоформ (каст, S)	19415	19252	163
S/U	0,263	0,261	0,002
Количество словоформ с n_i=1	13288	13112	176
Пойнтер-точка *R*	6127	6140	-13
Ранговое среднее	3014,4	2971,9	42,5
Индекс разнообразия	0,891	0,880	0,011

В этом случае данные получились более точные, так как в обработку не входило использование файнридера. Дореформенный текст был переведен на новую орфографию с помощью надстроек Word2007(RussianSpelling). Программа переводила слова со старой орфографии на новую в соответствии с внесенными в нее правилами. Следует заметить, что при таком способе обработки текстов появилась небольшая ошибка из-за автоматической работы программы, пропускавшей некоторые устаревшие формы слов. Например, слово «возстановлять» было переведено по правилам изменения написания приставок как «восстановлять», хотя такого слова в современном русском языке нет, вместо него следует употреблять «восстанавливать». Далее оба массива были обработаны с помощью программы URS.

Таблица 4. Характеристики массива по «Правилам…» Я.К.Грота в до- и пореформенной орфографии.

Характеристика	X_1	X_2	$X_1 - X_2$
Количество словоупотреблений (U)	18513	18511	2
Количество словоформ (каст, S)	6216	6098	118
S/U	0,335	0,329	0,006
Количество словоформ с n_i=1	4413	4303	110
Пойнтер-точка *R*	1803	1795	8
Ранговое среднее	1211	1170	41
Индекс разнообразия	0,889	0,881	0,008

Массив «Правил Орфографии» Я.К.Грота был также переведен с дореформенной на пореформенную орфографию с помощью надстройки Word2007, а затем по обоим массивам были получены данные, которые приведены в табл. 4.

Массивы из стихотворений отбирались параллельно в старой и в новой орфографии, после чего были получены количественные данные, которые приведены в табл. 5.

Таблица 5. Характеристики массива по поэтическим текстам в до- и пореформенной орфографии.

Характеристика	X_1	X_2	$X_1 - X_2$
Количество словоупотреблений (U)	22997	23115	-118
Количество словоформ (каст, S)	9025	8977	48
S/U	0,392	0,388	0,004
Количество словоформ с n_i=1	6705	6605	100
Пойнтер-точка *R*	2320	2372	-52
Ранговое среднее	1954,8	1932,4	22,4
Индекс разнообразия	0,911	0,903	0,008

Результаты по сводным массивам приведены в табл. 6.

Таблица 6. Характеристики сводного массива в до- и пореформенной орфографии.

Характеристика	X_1	X_2	$X_1 - X_2$
Количество словоупотреблений (U)	138847	139244	-397
Количество словоформ (каст, S)	34566	34122	444
S/U	0,248	0,245	0,003
Количество словоформ с n_i=1	30026	29562	464
Пойнтер-точка *R*	4541	4561	-20
Ранговое среднее	5279,9	5156,5	123,3
Индекс разнообразия	0,882	0,881	0,001

Данные, полученные по 10 изученным массивам, позволяют сделать следующие заключения.

Прежде всего, можно отметить, что во всех парах массивов наблюдается изменение числа словоупотреблений. Различие в количестве словоупотреблений определяется различием правил слитного и раздельного описания (было-бы, знаете-ли, какъ-бы, что-же, на-смерть, можетъ-быть, вовсе-бъ, ты-жъ, так-же поняли-ль, тому-жъ, то-есть, стало-быть, самолучшаго и пр., в 1956 – по-видимому, по-прежнему, по-пустому – через дефис вместо принятого раньше слитного написания, а вовремя – наоборот, слитно). При этом в двух случаях (рассказы Чехова и поэтические тексты) количество словоупотреблений возросло (на 1,17% и 0,51% соответственно), а в двух (Карлейль и Грот) практически не изменилось, что привело к увеличению после реформы числа словоупотреблений в сводном массиве (на 0,29%). Таким образом, изменение числа словоупотреблений при переходе с одной орфографии на другую является какой-то внутренней характеристикой текста и не может быть описано посредством универсального пересчетного коэффициента. Поэтому изучение изменения этого числа должно быть продолжено[1].

Число словоформ при переходе к новой орфографии уменьшилось по всем массивам.

Большее число словоформ в дореформенных текстах является следствием более низкого уровня омографии, определяемым большим числом используемых букв. Меньший уровень омографии проявляется также и в большем числе однократных словоформ в дореформенной орфографии. Меньшее значение пойнтер-точки *R* (во всех – в том числе, и в сводном, – массивах, кроме грамматики Грота) в дореформенной орфографии говорит о большем ценологическом разнообразии старых текстов. Индекс разнообразия по всем массивам (в двух орфографиях) практически не варьирует (0,88-0,91).

Таким образом, даже на небольшом материале видны систематические расхождения статистических распределений словоформ в одних и тем же текстах, записанных по правилам до- и пореформенной орфографии.

---

[1] Предварительные наблюдения наводят на мысль о том, что степень различия возрастает с увеличением сложности текста.

При этом первые двенадцать рангов распределения как в сводных массивах, так и в отдельных словарях дореформенной и современной орфографии представлены одними и теми же словоформами (табл.7) с примерно равными частотами. Некоторое различие в числе словоупотреблений связано с погрешностями обработки (напр., ошибки распознавания файнридера), а систематически меньшая их процентная доля в современной орфографии связана с большим количеством словоупотреблений в тексте. Тем не менее, это возрастание удельной доли первых рангов также «работает» на усиление асимметричности распределения, которое следует ожидать с точки зрения ценологического подхода. Лексически же начало словаря в обсуждаемом аспекте не представляет интереса т.к. представлено неизменяемыми словами, не имеющими омографов. Жирным выделены словоформы, отличающиеся по частоте.

Следующие ранги 13-35 представлены либо одними и теми же словами с несколько различающимися частотами (в 1-2 единицы, что, видимо, также определяется ошибками обработки), либо парами-тройками слов, меняющимися местами на 1-2 ранга. После этого каких-то выраженных характерных с лексикографической точки зрения особенностей распределений словоформ не обнаруживается.

Таким образом, речь идет о систематических различиях дореформенных и современных текстов. При этом *ранговое среднее* всех изданий в старой орфографии отличается от ранговых средних пореформенных изданий, причем разность эта как по критерию Стьюдента, так и по критерию Фишера является существенной. *Индекс разнообразия* и отношение S/U показывают, что все тексты до реформы орфографии были разнообразнее текстов после реформы. При этом эти два близкие по смыслу показателя (различаются только тем, что в индекс разнообразия введены логарифмы) указывают на одну и ту же тенденцию при рассмотрении всех пар массивов, однако дают разные оценки степени различия, причем выявить каких-либо тенденции, объясняющие различие этой степени, пока не удается.

Таблица 7. Самые частые словоформы.

Ранг	Словоформа	X_1		X_2	
		Частота	Доля (в %)	Частота	Доля (в %)
1	и	6580	0,04628	6580	0,04614
2	**в (въ)**	**4304**	**0,03027**	**4306**	**0,03029**
3	не	2649	0,01863	2649	0,01858
4	**на**	**1650**	**0,01160**	**1656**	**0,01161**
5	я	1568	0,01102	1568	0,01100
6	**что**	**1542**	**0,01084**	**1552**	**0,01088**
7	**с (съ)**	**1437**	**0,01011**	**1440**	**0,01010**
8	**а**	**1117**	**0,00786**	**1119**	**0,00785**
9	он (онъ)	1032	0,00726	1032	0,00724

10	**как (какъ)**	**924**	**0,00650**	**925**	**0,00649**
11	то	897	0,00631	892	0,00626
12	по	841	0,00592	841	0,00590

Общий вид распределения частот словоформ в массиве Чехова представлен на Рис. 1.

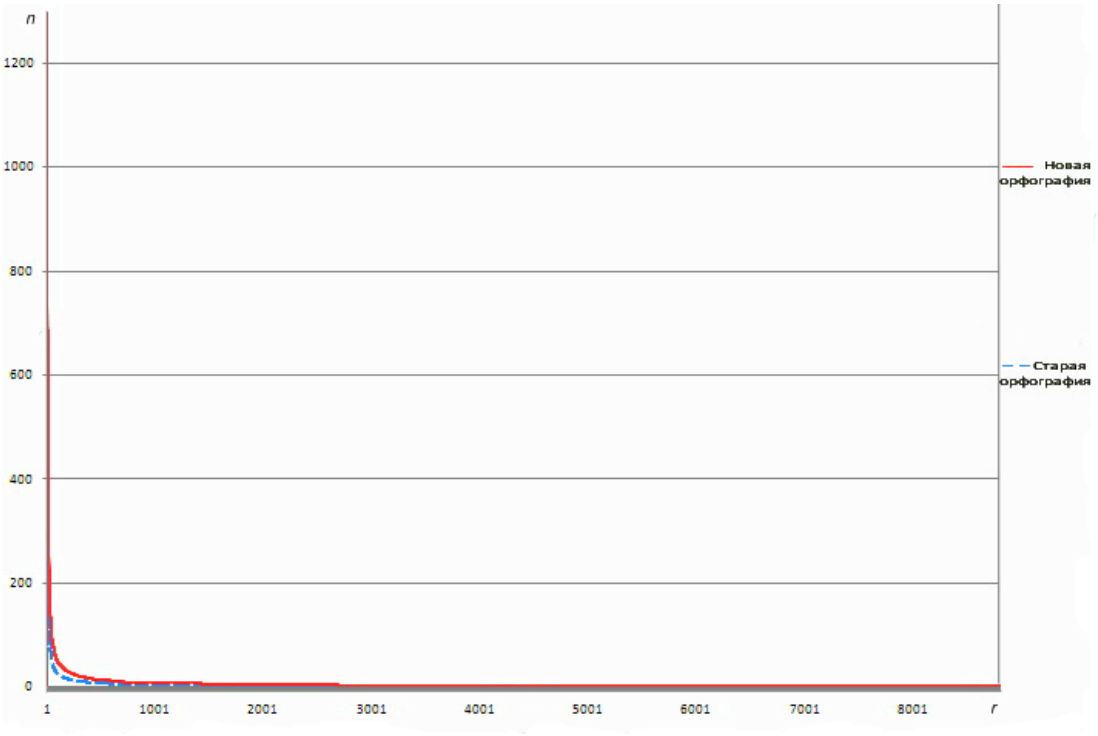


Рис. 1. Общий вид распределения частот словоформ в массиве Чехова.

На графике (Рис. 1) видно, что речь идет о распределениях практически тождественных в области первых рангов (при упоминавшейся несколько большей частоте в дореволюционной орфографии). При этом четко выражено превышение частоты в новой орфографии на значительном участке средней части распределения (примерно от 30 до 450 ранга – см. Рис. 2), а также более длинный хвост распределения в дореволюционной орфографии, что в целом также говорит о большем ценологическом разнообразии дореформенного текста. Полученный вывод может быть подтвержден и традиционными методами статистики.

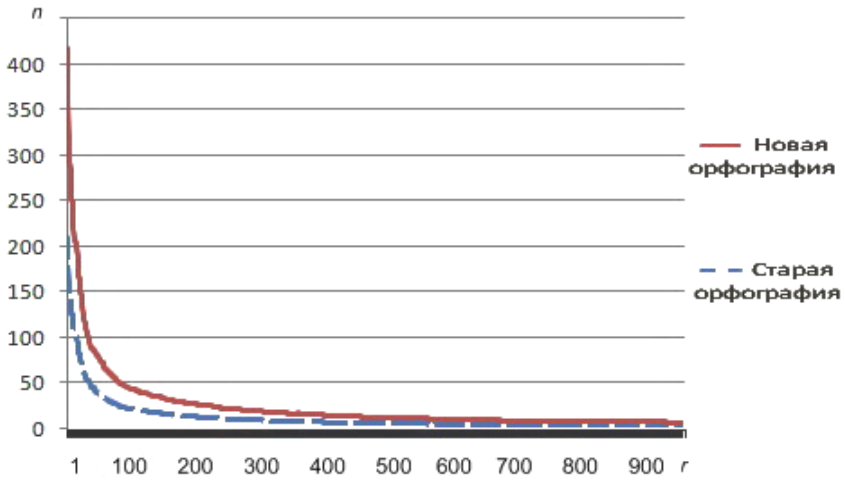


Рис. 2 . Средняя часть распределения частот словоформ в массиве Чехова.

Более внимательное изучение отдельных участков распределения (Рис. 2) показывает, что во всех исследованных массивах в центральной части распределения численность данного ранга в тексте на современной орфографии выше численности в орфографии дореформенной. Еще более детальное изучение материала (Плахотя, 2009, с. 49, Рис. 2, Плахотя, Чебанов, 2009, с. 227-228, Рис. 3-8) показывает, что первое небольшое превышение приходится на середину рангов четвертого десятка, второе, более заметное, на середину пятого десятка, затем на седьмой-десятый десятки (с максимальным расхождением на седьмом-восьмом) и далее численность ранга в тексте на современной орфографии никогда не бывает меньше, чем на дореформенной (вплоть до рангов, соответствующих уникальным словоформам, число которых, однако, в тексте с современной орфографией меньше). Отмеченное обстоятельство позволяет говорить, что рассматриваемые распределения достоверно различаются по *G*-критерию знаков традиционной статистики.

Отдельным направлением дальнейших исследований является изучение влияния реформы на рифму.

Для получения рекогносцировочных данных по этой проблеме были просмотрены стихотворения различных авторов (всего 9598 строк), написанные на старой орфографии. В них искались слова и окончания слов в конце строк, которые после реформы должны изменить звучание, что может привести к нарушению рифмы. Затем были найдены современные издания этих стихотворений с исправленными тем или иным способом словами для сохранения рифмы. Всего было обнаружено около 30 строк (0,31%), в которых было замечено нарушение рифмы после реформы орфографии, например:

Давыдов «Гр. П. А. Строганову».

Потомка бѣднаго ты пожалѣй **Батыя**,

И за чекмень прими его стихи **дурныя**!

Современное издание:

Потомка бедного ты пожалей **Батыя**
И за чекмень прими его стихи **дурные!**

Таким образом, в пореформенных изданиях изменение окончаний привело к нарушению рифмы (Батыя - дурные). Одни издания перепечатывали стихотворения, меняя слова по предписанным правилам реформы. Другие же, наоборот, пытаясь сохранить авторскую редакцию и не нарушить рифму, изъяли из письма лишь запрещенные буквы, оставив окончания автора (вышине – и оне).

Очевидно, что столь низкая частота (0,3%) нарушения рифмы не может заметно повлиять на культуру восприятия русского стиха в целом, однако, такие изменения могут безнадежно испортить отдельные стихотворения, превратив их в архаику, а, кроме того, привести к недооцениваемому смещению оценки отдельных лингвостатистических показателей.

Б) Ценологическая детализация

Для характеристики различия распределений была введена *мера различия,* равная отношению частот i-го ранга словника текста в новой орфографии к частоте того же ранга в старой $K_i=n_{i\ НОВ}/n_{i\ СТАР.}$ График меры различия для первых 400 рангов представлен на Рис. 3, на котором видно, что в подавляющем числе случаев $K_i \geq 1$. Значения $K_i <1$ появляются в основном (за исключением первой сотни рангов) на участках, в которых частота в тексте на новой орфографии поменялась на более низком ранге, чем в тексте на старой орфографии и принимает значение $K_i \geq 1$, когда в тексте на старой орфографии происходит уменьшение частоты.

Следует заметить, что введенная мера различия ведет себя сходным образом на всех изученных массивах.

На первом участке (несколько более 100 для Чехова, около 80 для Карлейля, около 80-140 у Грота, около 50-120 для стихов и около 80 в сводном массиве) имеют место резкие скачки значения K_i. При этом значения K_i могут быть как выше, так и ниже 1 (за исключением массива по Чехову, для которого K_i почти всегда больше единицы). Затем просматривается зона линейного роста K_i (для Чехова она намечается с первых рангов и четко видна примерно с 120 по 300 ранг, у Карлейля видна примерно с 200 по 300 ранги, хотя легко ее продолжить влево примерно до 100 ранга, у Грота – с 30 по 240 ранг, однако, с провалами на 100-170, примерно с 40 по 210 для стихов и с 70 по 400 для сводного массива).

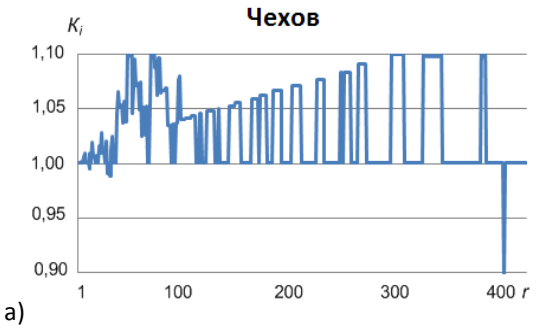
Далее возникает впечатление, что в трех массивах по Чехову, Гроту и стихам достигается максимум (при K_i =1,10 для Чехова r=290÷380, для Грота r=233÷235, для стихов r=300÷400). Для сводного массива и массива Карлейля такого максимума не наблюдается, хотя если аппроксимировать его достижение, то можно ожидать в обоих случаях при K_i =1,05. При этом у всех массивов на интервале r=300÷400 появляются отдельные значения K_i <1.

Такое поведение K_i интерпретируемо. Первая зона резких скачков относится к высокочастотной лексике, свойства которой отличны от лексики основного массива, отвечающего зоне линейного роста K_i, конец которой соответствует зоне максимальных различий частот одного ранга в распределение лексики до- и пореформенных текстов. Последняя же зона – это зона низкочастотной лексики, несовпадение скорости падения численности которой и дает K_i <1.

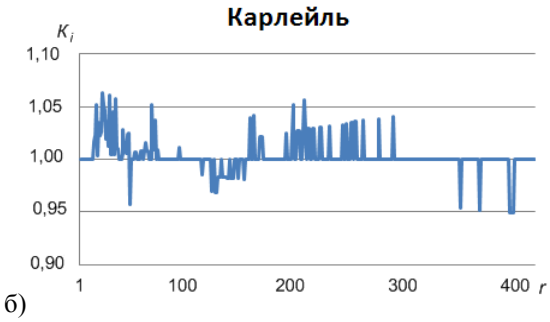
Воспроизводимость (хотя и не на большом числе массивов) характера различий текстов в до- и пореформенной орфографии и их интерпретируемость является еще одним свидетельством систематического характера различий текстов в двух формах записи.

Не менее интересно несходство поведения введенной меры различия для различных массивов.

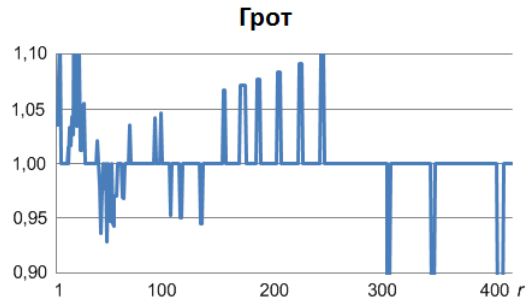
Во-первых, выше уже отмечалось, что качественно сходные особенности различий частот приходятся на разные ранги.



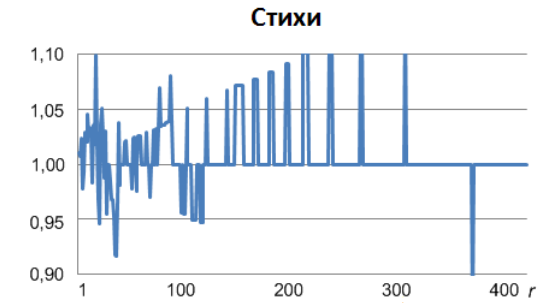
а)



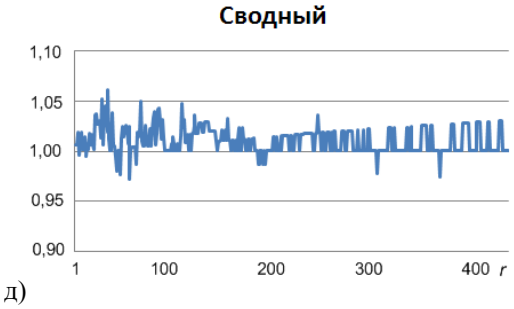
б)



в)



г)



д)

Рис.3. Распределение меры различия по рангам.

Во-вторых, есть различия в пределах варьирования K_i и значении его медианы. Так, массивы Карлейля и сводный характеризуются тем, что значения K_i почти полностью укладываются в интервал 0,95÷1,05 при значении медианы около 1,00 и в подавляющем большинстве случаев K_i >1 для сводного массива. Для массива Чехова, Грота и стихов K_i варьирует в пределах 0,90÷1,10, причем для массива Чехова K_i <1 являются очевидными исключениями. Возможно, что это связано с объемом выборки[2] и при увеличении объема выборки выше некоторого порогового значения эта величина монотонно убывает.

В-третьих, графики а-д рис. 3 демонстрируют наличие контрастных различий функции $K_i=f(r)$ для разных массивов, что показывает перспективность изучения этой зависимости.

В-четвертных, можно дать качественную содержательную интерпретацию особенностям функции $K_i=f(r)$. Наибольшее число скачков и наибольший размах варьирования характерен для наименее однородных массивов – стихотворений разных авторов, которые написаны с установкой на нарочитую оригинальность[3], и грамматики Грота, в которой сочетается научная терминология с примерами из обыденного языка (если эти различия в полной мере не являются следствием «малости» этих массивов – ср. выше «во-вторых»).

Для того, что дать интегральную характеристику степени отличия до- и пореформенного распределения словоформ в каждом массиве введем усредненную величину K_i

$$K_{ср}= (\Sigma K_i)/i,$$ где i – число рангов в каждом массиве[4].

Тогда для изученных массивов имеем величины, приведенные в табл. 8.

Таблица 8. Усредненное различие частот распределений словоформ в до- и пореформенных текстах

Массив	Число рангов	ΣK_i	$K_{ср}= (\Sigma K_i)/i$
Чехов	8010	8809,21	1,100
Карлейль	19415	19260,14	0,992
Грот	6216	6092,92	0,980
Стихотворения	9025	9040,38	1,002
Сводный	34566	34957,62	1,011

[2] Об этом свидетельствует и то, что на совсем небольшом объеме (рассказ А.П.Чехова «Шуточка» – около 1250 словоупотреблений), оказывается что K_i достигает значения 2.

[3] Тем более, что в ней широко представлены поэты второго и даже третьего ряда (что определяется наличием машиночитаемых текстов в старой орфографии), для которых «оригинальничание» бывает важно само по себе.

[4] В данном исследовании за число рангов в этих расчетах принято число рангов в массивах на старой орфографии, а частота отсутствующих классов в массивах на новой орфографии принимается равной нулю.

Полученные результаты показывают, что усредненная, удельная величина различий варьируется для разных массивов. При этом сводный массив имеем промежуточное значение этой величины, т.е. здесь имеет место какое-то ее усреднение. При этом здесь опять видно сходство массивов Карлейля и Грота. Возможно, это связано со способом обработки материала по этим массивам. В целом же полученные результаты еще ждут интерпретации.

Так или иначе, изучения мер различия распределений до- и пореформенной орфографии представляется перспективным и должно быть продолжено.

Для дальнейшего изучения характера различий распределений, последние были представлены в нормированном виде. При этом все частоты были заменены их долями:

$f_i=n_i/\sum n_i$ для i=1÷k, где k – количество классов.

В результате кривые графиков распределений для старой и новой орфографии изменились, что видно на графике (Рис. 4), на котором показаны 1-325 ранги (такое же соотношение прослежено до 400-го ранга).

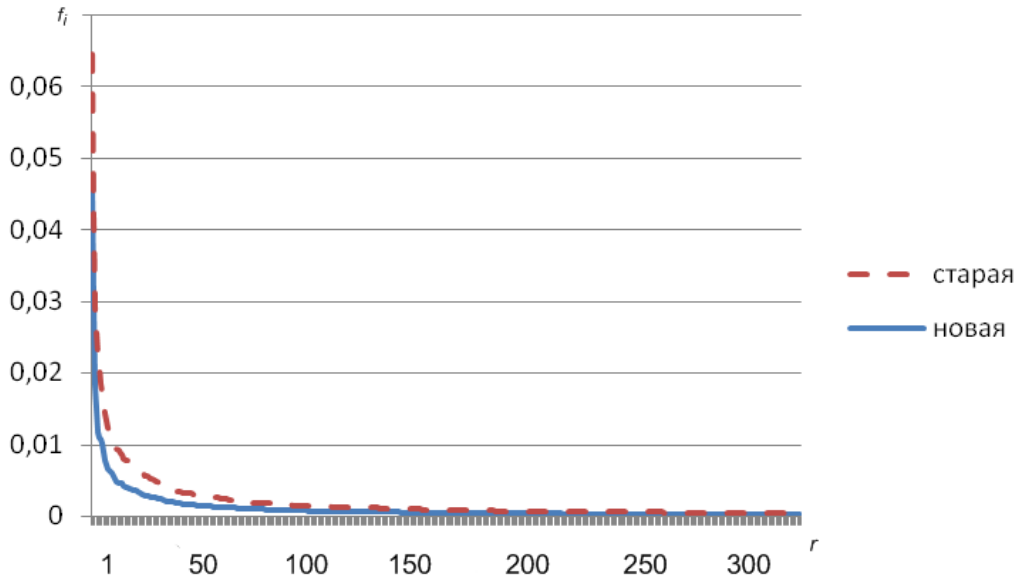


Рис.4. Распределение долей словоформ в массиве Чехова.

Представленные в таком виде данные стали пригодны для энтропийно-анэнтропийного анализа состава (лексического), разработанного Т.Г.Петровым (Петров, 2005, с.57).

Для сводного массива были получены следующие значения энтропии *H* и анэнтропии *A*:

Старая орфография

H =12,0924[5] *A*=16,42051

Новая орфография

H =12,0858 *A*=16,42129

[5] Принятое округление рекомендовано автором метода – Т.Г.Петрова. Функции не табулированы для таких значений степеней свободы.

Таким образом, *H* после реформы орфографии очень незначительно уменьшается (на 0,054%,), а *A* еще более незначительно, но увеличивается (на 0,005%). О значимости этих изменений можно будет судить по мере накопления данных и их осмысления[6].

Уменьшение шенноновской энтропии (*H* стала меньше) означает, что меньше стала неопределенность появления слов в тексте, или, что то же самое, – сложность текста (языка) снизилась. Это обстоятельство прямо говорит об упрощении русской письменности, причем рост *A* свидетельствует о том, что редкие элементы стали еще реже (исчезновение слов с ятями, ижицами и фитами). Направленность изменений энтропии и анэнтропии соответствует теоретическим ожиданиям. Еще больший интерес они могут представлять при продолжении данных исследований, например, при анализе статистических свойств текста в догротовской орфографии и с учетом реформы 1956 г.[7]. Исследования же энтропии и анэнтропии распределения словоформ по отдельным массивам пока не дают ясных результатов и требуют дальнейшего осмысления.

Полученные различия в числе словоформ и словоупотреблений несомненны в логике ценологического анализа, поскольку обнаружены все значимые в этом подходе различия.

Последнее обстоятельство позволяет поставить под вопрос корректность имеющихся лингвостатистических характеристик текстов начала XX, XIX и даже XVIII веков, в частности, использования этих характеристик для изучения стилистики, поэтики и эстетики текста и, в частности, художественного текста (в первую очередь, стихотворного).

В рамках принятой логики, можно дать ценологическую характеристику распределения букв в текстах. Это было сделано на массиве Карлейля, объемом около 450000 знаков (таб.9, Рис. 5-6).

Для ценологической характеристики представим эти распределения в ранговой форме (Рис. 6).

С ценологической точки зрения различия распределения букв как сходно, так и несколько отличается от различий распределения словоформ.

Прежде всего, в этом случае отсутствует тождественная высокочастотная зоны распределений. Вместо этого с 1 по 7 ранг сразу идет зона более низких частот в текстах на старой орфографии, а затем переходная зона с постепенно нарастающей разницей частот в пользу текстов на старой орфографии. Зона низких частот характеризуются (как и в случае словоформ) превышением частот в старой орфографии и более длинным «хвостом»,

---

[6] Замечание Т.Г.Петрова.

[7] В 1956 г. Академией наук СССР, Министерством высшего образования СССР и Министерством просвещения РСФСР утверждены «Правила русской орфографии и пунктуации», изменяющие некоторые орфографические нормы.

отвечающим большему числу букв в дореформенном русском алфавите. Такое положение свидетельствует о большей ценологической сложности буквенного состава дореформенного текста.

Таблица 9. Статистика букв массива Карлейля в до- и пореформенной орфографии (цветом выделены буквы, частоты которых различаются).

Ранги	Буквы	Новая орфография	Буквы	Старая орфография
1	О	45928	О	45084
2	Е	39159	Е	29770
3	И	29466	А	28101
4	Н	27666	Н	27666
5	А	27136	Т	26134
6	Т	26134	И	25054
7	С	21937	С	21588
8	В	17597	Ъ	21033
9	Р	16634	В	17597
10	Л	16448	Р	16634
11	М	14744	Л	16448
12	К	13387	М	14744
13	Д	12694	К	13387
14	Ы	9505	Д	12694
15	У	9266	Ы	9505
16	П	9223	У	9266
17	Б	8131	П	9223
18	Я	8027	Я	8915
19	Ь	7244	Ѣ	8535
20	Г	7108	Б	8131
21	З	6400	Ь	7234
22	Ч	6054	Г	7108
23	Й	5048	З	6749
24	Ж	4284	Ч	6054
25	Х	4069	Й	5048
26	Ш	2996	І	4397
27	Ю	2739	Ж	4284
28	Щ	1973	Х	4069
29	Э	1460	Ш	2997
30	Ц	1256	Ю	2739
31	Ф	1052	Щ	1972
32	Ё	156	Э	1460
33	Ъ	106	Ц	1256
34	Ѣ	0	Ф	1033
35	Ѳ	0	Ё	24
36	І	0	Ѳ	19
37	Ѵ	0	Ѵ	0

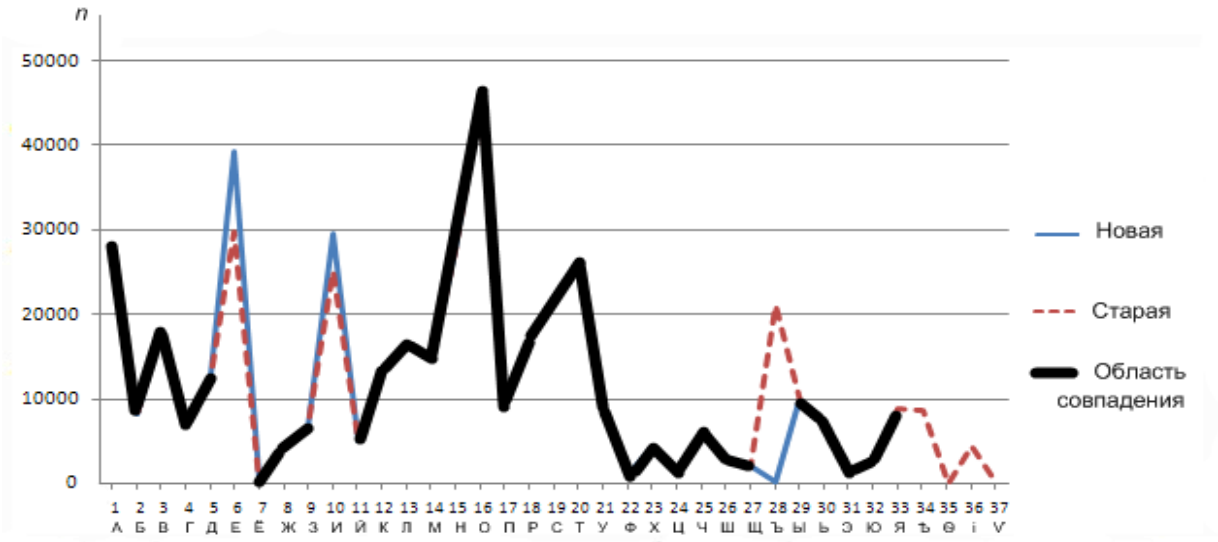


Рис. 5. Статистика букв массива Карлейля в до- и пореформенной орфографии

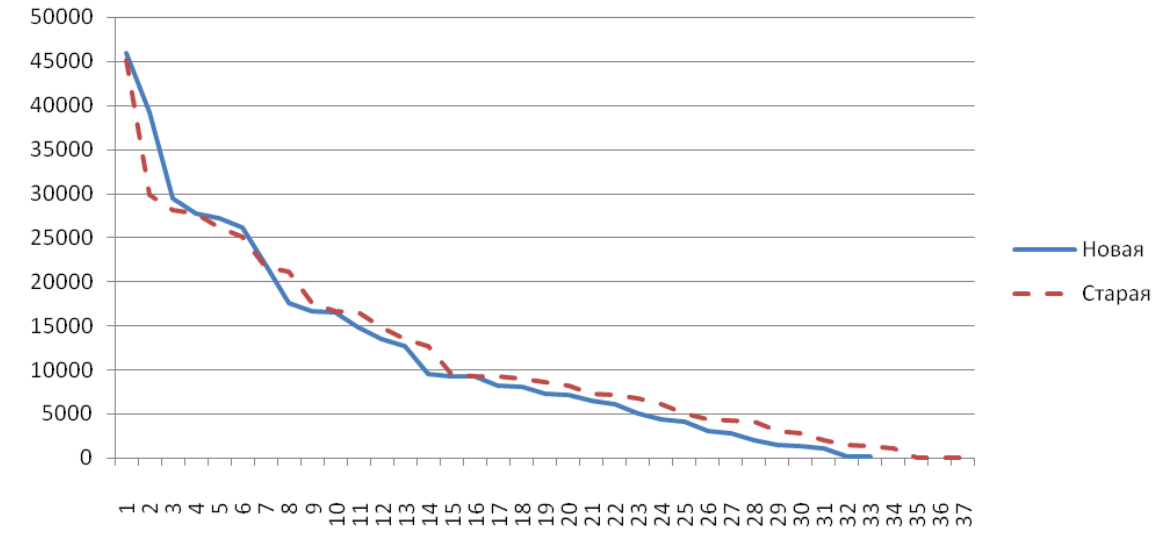


Рис.6. Распределние букв массива Карлейля в до- и пореформенной орфографии

Для распределения букв в этом массиве были получены значения энтропии и анэнтропии[8] (таб. 10).

Таблица 10. Энтропия и анэнтропия распределения букв массива Карлейля в до- и пореформенной орфографии.

	Старая	Новая	Разница
H	1,3925	1,3436	- 3,52%
A	1,6689	1,7856	+ 6,09%

Из таблицы видно, что проявляется та же тенденция, которая намечается для распределения словоформ, но, во-первых, она обнаруживается на значительно меньшей выборке, а, во-вторых, выражена значительно сильнее.

[8] Принятое округление рекомендовано автором метода – Т.Г.Петрова.

Содержательно такое положение дел вполне понятно – именно ожидание ощутимого полиграфического эффекта и было одним из мотивов проведения реформы, поэтому вполне понятно, что различия в распределении букв не вызывают сомнения. При этом имеет место совпадение фактического положения дел и теоретических ожиданий, предполагающих упрощение текста.

Примечательно при этом то, что это довольно сильное различие на уровне букв уже на уровне словоформ проявляется весьма слабо, а на уровне предложений можно ожидать полного отсутствия различий старой и новой орфографии. В этом проявляется целостность языка, в котором различия низших уровней структурной организации нивелируются организацией более высоких уровней.

Поскольку, как уже было сказано выше, с точки зрения ценологического подхода, тексты, подчиняющиеся *H*-распределению, трактуются как целостные, в частности, как написанные «на одном дыхании», то интересно рассмотреть не только массив по Чехову целиком, но посмотреть и особенностей распределений отдельных рассказов А.П.Чехова в старой и новой орфографии. При этом у отдельных рассказов можно ожидать большую степень «ценотичности», чем у всего массива.

С этой целью в каждом рассказе подсчитывались количество словоупотреблений, количество словоформ, число уникальных слов, соотношение словоформ/число словоупотреблений (S/U), пойнтер-точки R, а также меры различия (новая орфография/старая орфография). В результате для всех рассказов были получены данные, сходные с приводимыми ниже в таб. 11 и на графике 7 по рассказу «Шуточка».

Таблица 11. Характеристики рассказа «Шуточка» в до- и пореформенной орфографии.

Характеристика	X_1	X_2	$X_1 - X_2$
Количество словоупотреблений (U)	1244	1256	-12
Количество словоформ (каст, S)	627	619	8
S/U	0,50	0,49	0,01
Количество словоформ с f_i=1	479	465	14
Пойнтер-точка *R*	148	154	-6
Ранговое среднее	169,1	173,7	4,6
Индекс разнообразия	0,904	0,901	0,003

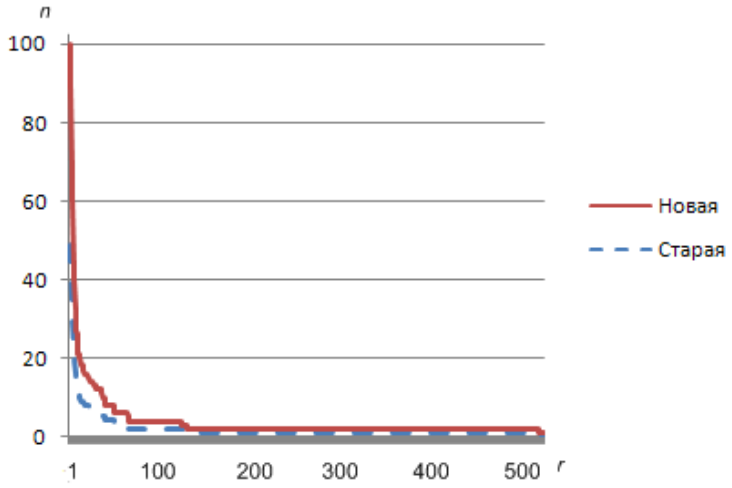


Рис. 7. Общий вид распределения частот словоформ в рассказе «Шуточка».

Меры различия $K_i=f_{i\ НОВ}/f_{i\ СТАР.}$ во всех рассказах качественно сходны между собой и, в частности, похожи на полученный для рассказа «Шуточка» (Рис. 8)

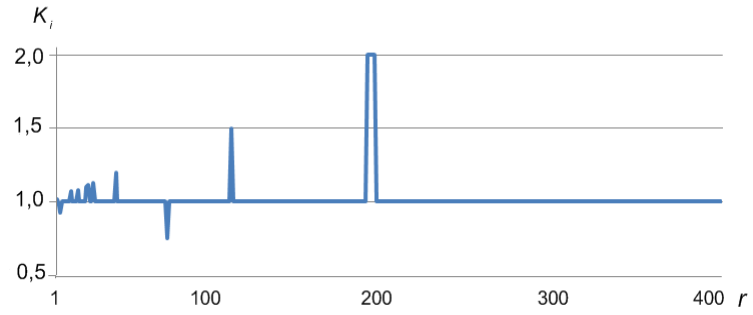


Рис.8. Распределение меры различия по рангам в рассказе «Шуточка».

Данные об изменении значения пойнтер-точки в рассказах Чехова приведены в табл. 12.

Полученные данные показывают, что имеем место та же, тенденция, которая обнаружена по всем массивам (кроме массива по Карлейлю), однако, различие оказывается значительно меньше (по-видимому, зависит от объема выборки), а в двух из 15 случаев и вовсе отсутствует.

После раздельного исследования рассказов А.П. Чехова можно сделать несколько выводов:

- Гипербола, описывающая распределение словоформ в новой орфографии располагается всегда выше гиперболы, описывающей распределение словоформ в старой.

Таблица 12. Значение пойнтер-точки в рассказах Чехова в до- и пореформенной орфографии.

Рассказ	$R_{стар}$	$R_{нов}$	$R_{стар} - R_{нов}$
Perpetuum mobile	250	256	-6
Беспокойный гость	217	230	-13
В Москве на Трубной площади	126	130	-4
В сарае	206	216	-10
Красавицы	296	300	-4
Лошадиная фамилия	105	108	-3
Мечты	295	309	-14
На Страстной неделе	162	171	-9
Нахлебники	221	226	-5
Следователь	180	185	-5
Сильные ощущения	166	166	0
Событие	163	170	-7
Старый дом	197	201	-4
Унтер Пришибеев	169	169	0
Шуточка	148	154	-6

- Рассмотренные характеристики этой пары распределений качественно сходны с характеристиками других пар, однако значения ряда из них и значения их разностей заметно меньше по абсолютной величине (количество словоупотреблений, количество словоформ, ноева каста, пойнтер-точка, ранговое среднее). Величина же S/U (как и во всех малых выборках) оказывается заметно выше, при том, что ее разность остается того же порядка, что и в других парах, а индекс разнообразия *I* как и его разность сопоставимы с таковыми для других пар.

- Мера различия $K_i=f_{iНОВ}/f_{iСТАР}$ показывает, что почти всегда частота i-го ранга распределения словоформ в новой орфографии более или равна таковой в распределении словоформ в старой орфографии.

- Систематическое отличие пойнтер-точек R показывает, что в каждом рассказе число маркированных (уникальных) слов текста в старой орфографии не меньше, чем в новой.

Заключение

На основании проделанной работы можно утверждать следующее.

Орфографическая реформа 1918 привела к незначительным (измеряемым первыми процентами или их долями) изменениям лингвостатистических характеристик одних и тех же текстов.

Оценка важности этих различий определяется принятыми статистической моделью и содержательно-лингвистической концепцией.

Во многих случаях обнаруженные различия до- и пореформенных текстов являются статистически достоверными по традиционным критериям (Стьюдента, Фишера, *G*-критерию знаков).

С ценологической точки зрения во всех парах до- и пореформенных текстов присутствуют сходные значимые различия, причем эти различия имеют содержательную интерпретацию.

Эти различия относятся к сводным массивам, четырем парам разножанровых массивов (рассказам А.П.Чехова, роману Т.Карлейля, грамматике Я.К.Грота, стихотворным текстам) и к отдельным рассказам А.П.Чехова, т.е. они воспроизводимы на выборках разного объема, разных жанров и разной степени однородности.

То обстоятельство, что разность характеристик сохраняется для выборок по крайней мере трех уровней (пара сводных массивов, четыре пары разножанровых массивов, отдельные рассказы Чехова), свидетельствует о статистической устойчивости этой величины, что из известных распределений характерно именно для негауссовых *H*-распределений (Шрейдер, Шаров, 1982).

Общей закономерностью для всех массивов является то, что и для словоформ и для букв (по массиву Карлейля) дореформенного текста спектр словоформ (букв) шире такового пореформенного, а значительная часть частот соответствующих рангов у дореформенного текста ниже, чем пореформенного (т.е. гипербола дореформенного текста лежит под гиперболой пореформенного – за исключением «хвоста», представленного ноевой касты). Этому соответствует несколько меньшее значение энтропии *H* и несколько большее значение анэнтропии *A*, что свидетельствует о ценологическом упрощении пореформенного текста (которое проявляется в снижении встречаемости редких слов и полном их исчезновения), уменьшении его целостности и выражается в увеличении доли неразрешимой омографии в пореформенном тексте.

Такое упрощение наряду с принижением этимологического принципа в орфографии, а также частичным (но небольшим) разрушением структуры стихотворного текста должно отражаться на упрощении речевой деятельности и мышления, что вполне соответствует духу времени, в которое эта реформа была завершена.

Многообразие изменений, вызываемых орфографической реформой 1918 г., дает основание рассматривать трансформацию текста с гротовской орфографии на

пореформенную не как нейтральное техническое преобразование, а как перевод[9], имеющий общекультурные последствия (т.е. это транскультурный перевод – Мартыненко, Чебанов, 1990).

Установленные факты (хотя малых, но систематических и статистически достоверных различий текстов в до- и пореформенной орфографии) должны приниматься во внимание при проведении лингвостатистических, лингвостилистических и стилеметрических исследований, причем характер этого внимания должен соответствовать сути осуществляемых разработок.

Так, если речь идет о поиске аналитического вида лингвостатистических распределений высокой степени приближения, то обнаруженные отличия должны учитываться или, по крайней мере, оговариваться в качестве принимаемых допущений при аппроксимации эмпирических распределений. В противном случае при достижении высокой степени точности статистические характеристики распределений будут иметь небольшую неучитываемую систематическую ошибку (безотносительно к тому, с какой целью исследуются эти распределения). Напротив, при проведении прицельных исследований отдельных языковых явлений (например, тех или иных стилеметрических характеристик) обнаруженные различия могут игнорироваться в том случае, если для осуществления этих исследований требуется значительно меньшая точность, чем величина обнаруженных различий. Оценить же психолингвистические и когнитологические последствия обнаруженных различий статистическими методами пока не представляется возможным.

Литература

Алексеев П. М. Квантитативная типология текста. Л., 1988.

Арапов М.В. Квантитативная лингвистика. М., 1988.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений // Научно-техническая информация. Сер. 2, 1975, №1.

Арапов М.В., Шрейдер Ю.А. Классификация и ранговые распределения // Научно-техническая информация. Сер. 2. 1977. №11-12.

Бодуэн-де-Куртенэ И.А. Об отношении русского письма к русскому языку. СПб., 1912.

[9] Таков был и вывод круглого стола «Принцип переводимости», прошедшего 10.10.2009 в рамках Тверской Международной герменевтической конференции «Понимание и рефлексия в коммуникации, культуре и образовании 2009».

Бычков В.Н. Построение, обоснование и интерпретация базисных лингвостатистических моделей. Дисс. ... к. филол. н. Л., 1986.

Валентинъ (Асмусъ) прот. О старой орѳографіи русскаго языка. Предисловіе къ брошюрѣ «Краткое пособіе по старой орѳографіи русскаго языка» // Русское зерцало, 1999 – http://russportal.ru/index.php?id=oldorth.asmus)

Гребенников А.О. Исследование устойчивости лексико-статистических характеристик текста. Дисс. … к.филол.н., СПб, 1998.

Декрет. О введении нового правописания // Известия В. Ц. И. К. 13 окт. 1918 г. № 223 (487) – http://lib.rus.ec/b/123799/read

Досекин Н.В. Обязательная неграмотность // Народоправство, 1917. Вып. 17.

Друговейко-Должанская С.В. Сколько букв в русской азбуке. СПб, 1998. (http://www.lingvotech.com/skolkobykv)

Иванов В.И. К вопросу об орфографической реформе // Вопросы жизни.1905. Вып. 9.

Ильин И.А. О русском правописании // Ильин И.А. Наши задачи, т. 2. Париж, 1956 – http://www.agencynews.ru/eliespecial/iljin_o_russkom_pravopisanii.pdf

Кудрин Б.И. Ценологические исследования распределений простых чисел (30-летие открытия). М.-Абакан, 2004.

Кудрин Б.И. Классика технических ценозов: Общая и прикладная ценология. Томск, 2006. (Ценологические исследования; вып. 31).

Кудрин Б.И. Мои семь отличий от Ципфа // Общая и прикладная ценология. 2007. № 4.

Кудрин Б.И., Лагуткин О.Е, Ошурков М.Г. Ценологический ранговый анализ в электрике. М., 2008.

Кучеров И.Я. Русское правописание и его реформирование. Б.г. – (http://slovnik.narod.ru/old/pravila/07.html, http://ocouchman.livejournal.com/591.html)

Левич А.П. Феноменология, применение и происхождение ранговых распределений в биоценозах и экологии как источник идей для техноценозов и ЭКОНОМИКИ Б.г. – http://www.chronos.msu.ru/RREPORTS/levich_fenomenologiya/levich_fenomenologiya.htm

Лопатин В. В. Русская орфография: задачи корректировки. М., Вып.5, 2001 (http://magazines.russ.ru/novyi_mi/2001/5/lopatin.html)

Мартыненко Г.Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других сложных системах // Структурная и прикладная лингвистика. Л., Вып.1.1978.

Мартыненко Г.Я., Чебанов С.В. Транскультурный перевод в специальной коммуникации. // Совещание-семинар “Этнопсихолингвистические аспекты речевого общения” Часть 1. М.-Самарканд, 1990.

Мефодий арх. К вопросу о реформе русского правописания. СПб, 1905.

Мечковская Н.Б. Язык и религия, М., 1998 (www.psylib.org.ua/books/mechk01/txt28)

Налимов В.В. Непрерывность против дискретности в языке и мышлении. Тбилиси, 1978.

Орлов Ю.К. О статистической структуре сообщений, оптимальных для человеческого восприятия // Научно-техническая информация. Сер. 2. 1970. №8

Орлов Ю.К. Динамика ранговых распределений и проблемы статистики редких событий // Математическое описание ценозов и закономерности технетики. Вып.1-2. "Ценологические исследования". Абакан, 1996.

Плахотя В.В. Влияние орфографической реформы 1917-1918 гг. на лингвостатистические закономерности русских письменных текстов. Выпускн. квалиф. работа. БГТУ «Военмех им. Д.Ф.Устинова. СПб, 2009.

Плахотя В.В., Чебанов С.В. Лингвостатистические последствия реформы русской орфографии 1918 г. // Технетика и ценология: от теории к практике. Ценологические исследования. Вып. 35. М., МОИП МГУ – ТЕХНЕТИКА, 2009.

Реформа русской орфографии 1918 года. Б.г. – http://ru.wikipedia.org/wiki/Реформа_русской_орфографии_1918_года

Сакулин П.Н. Реформа русского правописания. Пг., 1917.

Сандомирская И.И. Дух буквы // Искусство кино,1991, Вып.6.

Сборникъ статей о русском національномъ правописаніи. Б.м. 2006 - http://www.biblio.nhat-nam.ru/Rus_gram/articles.pdf

Троицкий В.П. ПАРЕРГА И ПАРАЛИПОМЕНА (статья Вяч. Иванова «Наш язык» – публикация, комментарии и размышления). Б.г. – http://losevaf.narod.ru/nashyaz.htm

Тюрин С. Болевые точки на теле русскаго языка – яти и еры – больное место русской культуры // Ундервудъ, 2000 (http://zudesnik.livejournal.com/3188.html)

Успенский Л.В. Слово о словах. Л., 1974.

Чебанов С.В. Внутренние и внешние системы в теории классификации // Системные исследования. 1979. М., 1980.

Чернышев В.И. Упрощение русского правописания. СПб.,1912.

Шагинян М. Тяпъ да ляпъ (Письмо в редакцию) // Русская воля, 1917, 2 июля.

Шрейдер Ю.А. О возможности теоретического вывода статистических закономерностей текста // Проблемы передачи информации. Вып. 1. М.: Наука. 1967..

Шрейдер Ю.А. Ранговые распределения как системное свойство // Математическое описание ценозов и закономерности технетики. Вып.1-2 "Ценологические исследования". Абакан, 1996.

Шрейдер Ю.А., Шаров А.А. Системы и модели. М., 1982

Щерба Л.В. Избранные работы по русскому языку. М., 1957. (http://www.ruthenia.ru/apr/textes/sherba.htm)

Языков Д.И. Замечания о некоторых русских буквах. СПб, 1809 (http://stroev.boom.ru/stroev07.htm).

Herdan, G. Quantitative Linguistics. Berlin, Heidelberg, London. 1964

Zipf G.K. Selected Studies of the Principle of Relative Frequency in Language. Cambridge (Mass.). 1932.