Методика поиска степени родства языков по чередованию гласных и согласных букв в письменных источниках.

Сведения об авторах. Инженер-программист НТЦ Модуль, Филатов О.В., г. Москва.

Аннотация. Разговорный язык характеризуется числом L. Все разговорные языки упорядочиваются по числам L вдоль единой оси координат. Родство языков определяется близостью величин характеризующих их L чисел. Получение L чисел производится посредством правил потоковой теории, путём конвертации текста в бинарную последовательность, с её последующей программной обработкой. Выдвигается гипотеза о связи возраста языка с его удалением от начала L оси.

Ключевые слова. Квантитативная лингвистика, бинарная последовательность, потоковая теория, составное событие, разговорный язык, родство языков, выпадение монеты.

Работа относится к квантитативной лингвистике. В ней описывается способ получения числовых оценок языков. Наличие числовых оценок языков позволяет ввести цифровую шкалу, и разместить на ней числовые величины характеризующие языки. Близость на шкале двух числовых величин означает родство символизируемых этими величинами языков.

Математическим аппаратом для ранжирования языков явилась недавно разработанная «Потоковая теория» случайных событий [1,2].

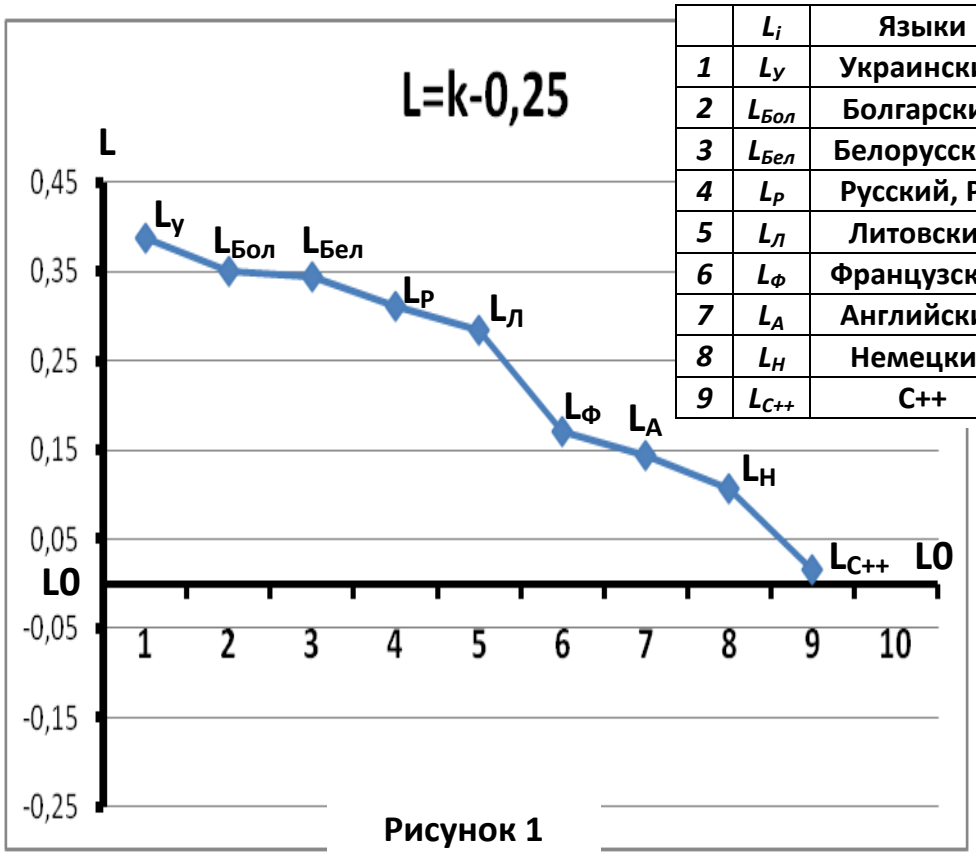
В работе предполагается, что каждый разговорной язык можно охарактеризовать одним числом. Обозначим это число от английского слова Language (язык, речь) буквой L. Числа L_i отражают удаление каждого

разговорного языка от природной (естественной) кривой распределения случайных составных событий L_0 (L_0 выводится из потоковой теории).

Основанием применения математического аппарата из «Потоковой теории» явилась похожесть чередования гласных и согласных букв в словах с выпадением случайных событий в потоковой последовательности. И желание автора узнать насколько разговорные языки отличаются от случайного набора бит.

В работе сравниваются цепочки из гласных и согласных букв различных языков, с математической кривой случайных длин составных событий из бинарной потоковой последовательности. Полученные числа L_i, характеризующие языки, и график, построенный из них, представлены на рисунке 1.

Вертикальная ось рисунка является языковой L-осью. Горизонтальная вспомогательная L0-ось обозначает природное (естественное) распределение случайных длин составных событий (первой моды [1,2]) при подбрасывании монеты. С введением природного нулевого L0 уровня языки получили точку отсчёта и свои положения на L-оси.



	L_i	Языки	L=k-0,25
1	$L_У$	Украинский	0,387
2	$L_{Бол}$	Болгарский	0,351
3	$L_{Бел}$	Белорусский	0,345
4	$L_Р$	Русский, РФ	0,320
5	$L_Л$	Литовский	0,285
6	$L_Ф$	Французский	0,171
7	L_A	Английский	0,143
8	L_H	Немецкий	0,106
9	L_{C++}	C++	0,015

Рисунок 1

Для этой работы были исследованы девять языков. Среди них четыре славянских языка и произошедший от общего протославянского языка литовский язык.

Числа L характеризующие славянские и другие европейские языки были бы перемешаны друг с другом, если бы L числа не имели бы связи с языками. Но обработка сгруппировала славянские языки одной группой L чисел, и поместила рядом с ними L-число родственного им литовского языка.

В общий ряд языков был включён и язык программирования C++, точнее его BDS 6.0 диалект.

Следует отметить, что все обработанные языки разместились на L-оси в области положительных значений.

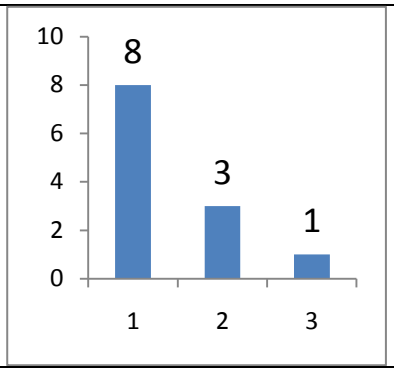
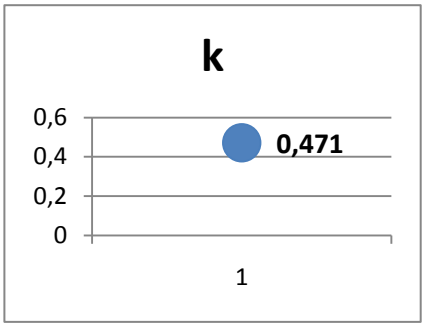
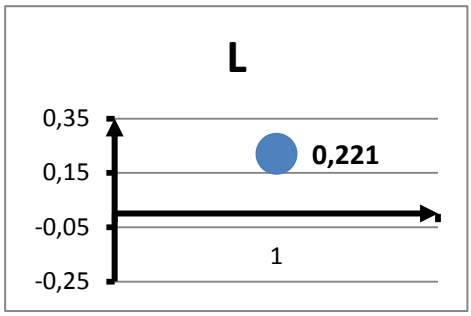
Трансформация языкового текста в потоковую последовательность.

Рассмотрим предлагаемую методику поиска степени родства языков на примере одной короткой фразы: «Какая милая орхидея!».

Превращение этой фразы в материал для компьютерного анализа продемонстрировано в таблице 1.

Таблица 1

	Информационные состояния	Объекты информации	Количество
1	Исходный текст	Какая милая орхидея!	20
2	Знаковая очистка	Какаямилаяорхидея	17
3	Бинарная интерпретация	10100101000110100	17
4	«1», «0»	101**101*****01**	8
5	«11», «00»	***00******11**00	3
6	«111», «000»	********000******	1

7	Результат - гистограмма длин составных событий		
8	Получение k; k=8/17= 0,471		
9	Получение L; L=k-0,25=0,471-0,25=0,221		
10	Размещение k или L в таблице или на графике.		
			

Разбор действий в таблице 1.

В строке 1 содержится фраза, которая подвергнется компьютерной обработке. Общее число букв, пробелов и знаков препинания в этой фразе равно 20.

В строке 2 исходный текст лишился всех знаков препинания и пробелов. Все слова фразы сжаты вместе, в одну длинную буквенную последовательность.

В строке 3 буквенная последовательность превращена в нули и единицы бинарной последовательности. Нулями стали все гласные буквы, а единицами стали все согласные буквы. Полученная бинарная последовательность готова к анализу. Образно говоря – бинарная последовательность сохранила в себе отпечаток фразы: «Какая милая орхидея». А большой кусок текста, по сути, являющийся частью разговорного языка, превращённый в бинарную последовательность, оставит в ней отпечаток этого языка.

В строках 4, 5, 6 показан анализ бинарной последовательности. Во время анализа были найдены составные события из нулей и единиц.

Так в строке 4 показаны найденные не повторяющиеся дважды события. То есть, если монета выпала орлом, то следующее выпадение произошло решкой. И более того, перед выпавшим орлом было так же выпадение решки. Таких событий найдено восемь.

В строке 5 показаны найденные двойные события. То есть, когда монета два раза подряд выпадала решкой (орлом). Таких событий найдено три.

В строке 6 показано единственное найденное событие тройного выпадения подряд монеты одной стороной.

В строке 7 приведена гистограмма из найденных в строках 4,5,6 событий. Собственно для нахождения рассматриваемых в этой статье L параметров используется только первый столбец гистограмм (в примере в нём восемь событий). Первый столбец в потоковой теории относится к первой моде элементаров (элов) [1,2]. И L параметры рассчитываются именно из элов потоковой теории.

В строке 8, число событий (элов) обладающих единичной длиной (взятых из строки 4) делится на длину бинарной последовательности (найденную в строке 3). Полученный результат уже может характеризовать фразу (язык). Этот результат помещён на график ‘k’ в строке 10.

В строке 9, для облегчения сравнения полученной величины k – характеризующей фразу (язык) с естественным (природным) уровнем производится приравнивание этого уровня (равного 0,25) к нулю. Это производится вычитанием из k значения 0,25. В результате этого действия получается число L. Оси координат и точка графика соответствующая L показаны в строке 10 на графике ‘L’.

Несколько слов о рабочей гипотезе

В начале работ автор придерживался рабочей гипотезы, что более древней язык должен быть менее случайным, более устоявшимся в тысячелетиях (столетиях). И поэтому он должен быть дальше удалён от природной случайной оси L0 (рисунок 1).

И наоборот, чем младше по возрасту язык, тем ближе он расположен к случайной оси L0. И, в рамках рабочей гипотезы и обработанных языков, украинский язык является самым древним, а “разговаривающие” на языке C++ программисты являются представителями самого молодого языка.

Так же сделаю предположение, что логика и мышление носителей языка, который окажется ниже случайной оси L0, в отрицательной области, должны отличатся от логики и мышления носителей языков расположенных выше случайной оси L0.

Но название для этой статьи было дано менее амбициозное, чем рабочая гипотеза, и более интуитивно понятное.

Отношение статьи к лингвистике.

Из статьи понятно, что её автор демонстрирует новый инструментарий для профессиональных лингвистов. Себя к таковым (профессиональным лингвистам) автор не относит. Статья написана в качестве ознакомительного примера по применению возможностей потоковой теории в качестве лингвистического инструмента. Хотя автор и смеет мечтать о том, что полученные лингвистические результаты будут рассматриваться как предварительные и черновые научные достижения. Но для получения более

качественных, итоговых, результатов у автора отсутствуют профессиональные лингвистические знания.

Поэтому статья максимально облегчена, не перечислены имена писателей и названия их произведений, взятые в качестве языковых источников. Так как тексты для подобных исследований должны отбираться по многим параметрами. Отобранные тексты должны гарантировать репрезентативность и чистоту языка. А этого автор статьи обеспечить не мог.

Сложность подбора исходных текстов, и только по Русскому языку, приводится на примере таблицы 2.

Таблица 2

500000 эл	1	2	3	4	5
Русский язык	Rowling перевод 273529	Пушкин поэзия 276604	Достоевский 282487	Толстой 283094	Шолохов 288915
L_P	0,297058	0,303208	0,314974	0,316188	0,32783
Среднее			0,319664		

В таблице 2 приведены расчеты чисел L_i, отражающих удаление русского языка от природной оси L_0, для нескольких авторов.

Параметр L рассчитанный по поэзии Пушкина ощутимо отличается от аналогичных параметров для прозы классиков. Причём, L(Пушкина) настолько же удалён от L(Достоевского), Насколько L(Достоевского) удалён от L(Шолохова).

Бросается в глаза, что перевод на русский язык книги «Гари Поттер и орден Феникса» обладает такой же величиной параметра L(Rowling)= 0,297058, как и величина стихотворного параметра L(Пушкина)= 0,303208.

Наличие L параметров с различными величинами для разных писателей говорит о том, что график 1 должен образовываться для представления разных языков не точками, а протяжёнными областями. Так, исходя из таблицы 2, область Русского языка на графике 1 нужно представить в виде протяжённой области:

0,297058 <= L_P <=0,32783

И так по каждому из языков. Но по причинам, озвученным выше в этом подразделе, автор решил не усложнять. Ведь задача этой статьи только предложить профессиональным лингвистам новый рабочий инструмент, основанный на потоковой теории.

Несколько слов о «Потоковой теории»

Поскольку представленные в статье результаты были получены на основе потоковой теории, то вполне логично молвить о ней слово.

Так как потоковая теория изучает физический процесс случайного выпадения монеты (или эквивалентные ему физические процессы), то она относится к физическим, вернее к физико-математическим теориям.

Если записывать результаты подбрасываний монеты через нули и единицы, то получается хаос, подобный этому: «10110010100001001000111011110101 1…». Но, в потоковой теории показано, что пропорции этого хаоса можно описать одной простой формулой.

Вот пример, который нельзя решить средствами комбинаторики. Пусть надо найти сколько раз выпадут подчёркнутые комбинации из четырёх нулей (0000) в двадцати миллионах подбрасываний монеты.

Ответ: $\frac{N}{2^{k+2}} = \frac{20000000}{2^{4+2}} = 312\ \ 500$ раз

То есть, весь хаос выпадений монет описывается одной короткой формулой, которая распределяет выпадающие составные события типа: «111», «000», «1111», «0000» и т.д. по пропорциям, в зависимости от числа бросков монеты.

Потоковая теория не может предсказывать выпадения стороны монеты с вероятностью иной, чем привычная вероятность 0,5. Но зато потоковая теория может предсказывать на основе выпавших состояний монеты

выпадения будущих составных событий. Что является принципиально новым моментом в теории бинарных последовательностей с вероятностью выпадения событий 0,5. То есть, исследовав уже выпавшие результаты подбрасывания монеты можно отстраняться или участвовать в будущих их выпадениях, что даёт получение разных пропорций в фиксируемых потоках составных событий.

Коротко говоря, потоковая теория анализирует поток выпадающих событий подбрасываемой монеты и принимает решения делать ставку или нет.

Библиографический список

1. Филатов О. В., Филатов И.О., Макеева Л.Л. и др. «Потоковая теория: из сайта в книгу». М.: Век информации, 2014. С.200.
2. Филатов О. В., Филатов И.О., Статья «О закономерностях структуры бинарной последовательности», «Журнал научных публикаций аспирантов и докторантов», № 5, 2014.