ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 16 за 2024 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

М.Ю. Кислицына, Ю.Н. Орлов

Распределение порядковых частот согласных букв как инвариант языковой группы

Статья доступна по лицензии
Creative Commons Attribution 4.0 International

Рекомендуемая форма библиографической ссылки: Кислицына М.Ю., Орлов Ю.Н. Распределение порядковых частот согласных букв как инвариант языковой группы // Препринты ИПМ им. М.В.Келдыша. 2024. № 16. 18 с. https://doi.org/10.20948/prepr-2024-16
https://library.keldysh.ru/preprint.asp?id=2024-16

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

М.Ю. Кислицына, Ю.Н. Орлов

Распределение порядковых частот согласных букв как инвариант языковой группы

Москва – 2024

Кислицына М.Ю., Орлов Ю.Н.

Распределение порядковых частот согласных букв как инвариант языковой группы

Собрана статистика распределения частот согласных букв в основных современных языках индоевропейской семьи. Изучались распределения упорядоченных по убыванию частот, построенные на основе анализа литературных текстов длиной порядка 1 млн знаков. Показано, что можно ввести инвариант языковых групп – германской, романской, славянской и балтийской – как расстояние между элементами группы в норме L1. Определено пороговое расстояние, при котором языки объединяются в группы как полносвязные подграфы, оно равно 0,14. Показано также, что структуры графа ближних и дальних соседей соответствует модели зависимых случайных величин.

Ключевые слова: машинная классификация, предобработка текстов, распределение упорядоченных частот, граф ближайших соседей

Kislitsyna M.Yu., Orlov Yu.N.

The distribution of ordinal frequencies of consonants as an invariant of a language group

The statistics of the frequency distribution of consonant letters in the main modern languages of the Indo-European family are collected. The distributions of descending frequencies were studied, based on the analysis of literary texts with a length of about 1 million characters. It is shown that it is possible to introduce an invariant of language groups – Germanic, Romance, Slavic and Baltic – as the distance between the elements of the group in the L1 norm. The threshold distance at which languages are grouped as fully connected subgraphs is 0.14. It is also shown that the structures of the graph of near and far neighbors correspond to the model of dependent random variables.

Keywords: machine classification, text preprocessing, ordered frequencies distribution, nearest neighbor graph

Содержание

Введение

Машинная обработка текстов на естественных языках (NLP, Natural Language Processing) представляет одно из важных направлений применения технологий искусственного интеллекта в задачах классификации текстов и массивов документов [1-3]. Существует большое количество компьютерных программ для такой обработки. Все они используют методы статистики для автоматического определения языка, жанра или тематики некоторого достаточно большого текста. В основном ищутся наиболее часто употребляемые слова или буквы.

В нашем исследовании мы будем рассматривать вопрос о том, существует ли вычисляемый инвариант, классифицирующий языки индоевропейской языковой семьи по языковым группам. Эта задача связана с автоматическим определением языка текста по частоте употребляемых в нем символов.

Ранее в работе [4] была предпринята попытка определить язык, на котором написан так называемый Манускрипт Войнича, для чего потребовалось провести анализ частотно упорядоченных символов в некоторых европейских языках на основе анализа соответствующих литературных текстов. Выяснилось, что алфавитное упорядочение букв не дает возможности построить достаточно точный индикатор языка, поскольку языки одной группы имеют весьма различные распределения. Напротив, если упорядочить символы по частоте встречаемости, то такое распределение оказывается более устойчивым.

Данные по частотам использования букв в разных языках приводятся во многих справочниках и аналитических работах. Однако эти данные довольно сильно различаются. Например, в книге [5] приведены следующие данные по частотам букв в некоторых языках (таблица 1), которую мы дополнили данными [6] для английского языка.

Из таблицы 1 следует, что данные по одним и тем же частотам для английского языка варьируются от 2% до 90%, что весьма неточно.

Заметим также, что данные [5] приведены с точностью до четвертого знака после запятой, а данные [6] – с точностью до пятого. Следовательно, декларируемая относительная точность самого часто встречаемого символа составляет в [5] примерно 10^{-4}, а в [6] – 10^{-5}. Поскольку это есть в то же время ширина доверительного интервала для оценки частоты, из стандартного критерия Стьюдента следует, что число данных для такой точности должно быть порядка 10^{10}. Для достижения такой точности надо проанализировать сто тысяч книг размером по 100 страниц, что кажется сомнительным, так как по порядку величины это вообще вся литература на определенном языке. Более достоверным представляется знание частоты с точностью до 10^{-2}, и то для текстов определенной тематической направленности, поскольку частотность зависит от того, какие именно тексты рассматриваются: по определенной профессиональной тематике, словари или художественная литература. Авторы, однако, не приводят список источников, обработка которых привела к тем или иным результатам. Это вполне оправданно, поскольку такой список сам займет не один десяток страниц и собственно научного интереса не представляет.

Таблица 1 – Частоты (%) употребления некоторых букв в иностранных языках

Буква алфавита	Французский язык	Немецкий язык	Английский язык		Испанский язык	Итальянский язык
			[5]	[6]		
A	7,68	5,52	8,167	7,96	12,90	11,12
B	0,80	1,56	1,492	1,60	1,03	1,07
C	3,32	2,94	2,782	2,84	4,42	4,11
D	3,60	4,91	4,253	4,01	4,67	3,54
E	17,76	19,18	12,702	12,86	14,15	11,63
F	1,06	1,96	2,228	2,62	0,70	1,15
G	1,10	3,60	2,015	1,99	1,00	1,73
H	0,64	5,02	6,094	5,39	0,91	0,83
I	7,23	8,21	6,966	7,77	7,01	12,04
J	0,19	0,16	0,153	0,16	0,24	-
K	-	1,33	0,772	0,41	-	-
L	5,89	3,48	4,025	3,51	5,52	5,95
M	2,72	1,69	2,406	2,43	2,55	2,65
N	7,61	10,20	6,749	7,51	6,20	7,68
O	5,34	2,14	7,507	6,62	8,84	8,92
P	3,24	0,54	1,929	1,81	3,26	2,66
Q	1,34	0,01	0,095	0,17	1,55	0,48
R	6,81	7,01	5,987	6,83	6,95	6,56
S	8,23	7,07	6,327	6,62	7,64	4,81
T	7,30	5,86	9,056	9,72	4,36	7,07
U	6,05	4,22	2,758	2,48	4,00	3,09
V	1,27	0,84	0,978	1,15	0,67	1,67
W	-	1,38	2,360	1,80	-	-
X	0,54	-	0,150	0,17	0,07	-
Y	0,21	-	1,974	1,52	1,05	-
Z	0,07	1,17	0,074	0,05	0,31	1,24

Из таблицы 1 также следует, что расстояния между распределениями в гистограммной норме составляют по данным [5]: французский–немецкий 0,38; немецкий–английский 0,49; английский–испанский 0,26; испанский–итальянский 0,35. То есть расстояние между языками германской группы (немецкий и английский) гораздо больше, чем между языками разных групп (германской и романской). Это означает, что алфавитное упорядочение не дает правильного представления о близости языковых групп.

Следовательно, наблюдение [4] о том, что близкими являются распределения упорядоченных частот встречаемости согласных символов, может дать возможность найти такое критическое расстояние между распределениями, ниже которого кластеризуются языки одной языковой группы. Мы предполагаем, что каждой языковой группе свойственно свое распределение ранжированных частот букв по убыванию. Тогда такую упорядоченность можно считать свойством лексикона.

Цель нашего исследования состоит в том, чтобы, изучив статистику частот употребления согласных букв в литературных текстах на языках индоевропейской языковой семьи, построить эталонные распределения частот для этих языков и сформулировать критерий объединения языков в группы на основе попарной близости этих эталонных распределений. Для построения распределений мы будем рассматривать только литературные произведения.

1. Методика исследования

Согласно справочнику [7], индоевропейская семья языков включает в себя двенадцать языковых групп. Поскольку в нашей работе мы будем анализировать современные тексты, написанные на латинице и кириллице, то из исследования исключаются группы так называемых мертвых языков и языков, на которых написаны древние тексты. Тогда рассматриваем следующие группы языков.

1. Балтийская (северобалтийская) группа содержит литовский и латышский языки.

2. Славянская группа языков подразделяется на три подгруппы: южнославянскую, западнославянскую и восточнославянскую. В этой группе мы рассматриваем языки: болгарский, македонский, сербский, хорватский (южная ветвь), чешский, словацкий и польский (западная ветвь), русский, белорусский и украинский (восточная ветвь). Однако поскольку хорватский язык использует латиницу, а все остальные языки южной подгруппы используют кириллицу, мы будем относить для удобства анализа хорватский язык к западной славянской группе.

3. Германская группа подразделяется на две подгруппы: северную и западную. Северная подгруппа содержит исландский, шведский, датский и норвежский языки. Западная подгруппа включает английский, немецкий и голландский языки.

4. Романская группа состоит из восточной и западной подгрупп. К восточной подгруппе относится румынский язык, а к западной – испанский, португальский, французский и итальянский языки.

Всего, таким образом, мы изучаем 24 европейских языка.

Для того чтобы частоты были оценены до второго знака после запятой, необходимо иметь тексты суммарной длины порядка 10^7 знаков. Сбор такого количества текстов представляет достаточно трудоемкую техническую задачу, так как требуется собрать около 50 крупных литературных произведений на 200-300 страниц (по 1000 знаков на страницу).

Далее надо провести определенную предподготовку текстов, поскольку в большинстве языков широко используются диакритические символы, указывающие на особенности произношения букв. Все такие символы мы удаляем, заменяя, например, «č» на «c» и т.д.

Пусть $f_i(j)$ – частота символа ранга j в i-м языке. Подчеркнем, что частоты нормированы на полное число согласных букв в тексте на данном языке. Ранжирование осуществляется по убыванию частоты. При этом на j-м месте в разных языках могут стоять разные буквы. Для нашего анализа важна именно последовательность убывающих частот.

Согласных букв после предобработки оказывается 20 для всех языков. Составляется матрица расстояний между распределениями частот:

$$\rho_{ik} = \sum_{j=1}^{20} \left| f_i(j) - f_k(j) \right| .$$

Далее проводится кластеризация языков. Рассматриваются два варианта. Первый – по взаимной близости всех эталонов, входящих в кластер. Для этого

требуется найти такое критическое значение расстояния ρ^* между эталонами, что для всех эталонов группы попарные расстояния меньше этого числа. Число ρ^* подбирается так, чтобы результат кластеризации был максимально приближен к априорно известным лингвистическим группам, а ошибка ложного включения была минимальна. Это важное замечание, поскольку можно взять $\rho^* = \max \rho_{ik}$ и получить полную кластеризацию всех языковых групп в одном кластере, что, очевидно, непродуктивно. Поэтому ищется минимальное значение ρ^* такое, что оно все же позволяет собрать в один класс большинство языков определенной группы или подгруппы.

Второй вариант – кластеризация методом ближайших соседей. Составляется направленный граф в соответствии с тем, какой эталон оказывается ближайшим к данному.

Полученные результаты мы также сравниваем с теми, которые получаются из анализа корреляций между распределениями частот и покажем, что подход с использованием близости эталонов в норме L1 существенно более точный.

2. Результаты расчетов эталонных распределений

На рис. 1-9 показаны распределения частот букв в различных языках. Частоты упорядочены по убыванию. Первый ранг – самый часто встречаемый символ, второй ранг – следующий за ним по частоте и т.д.

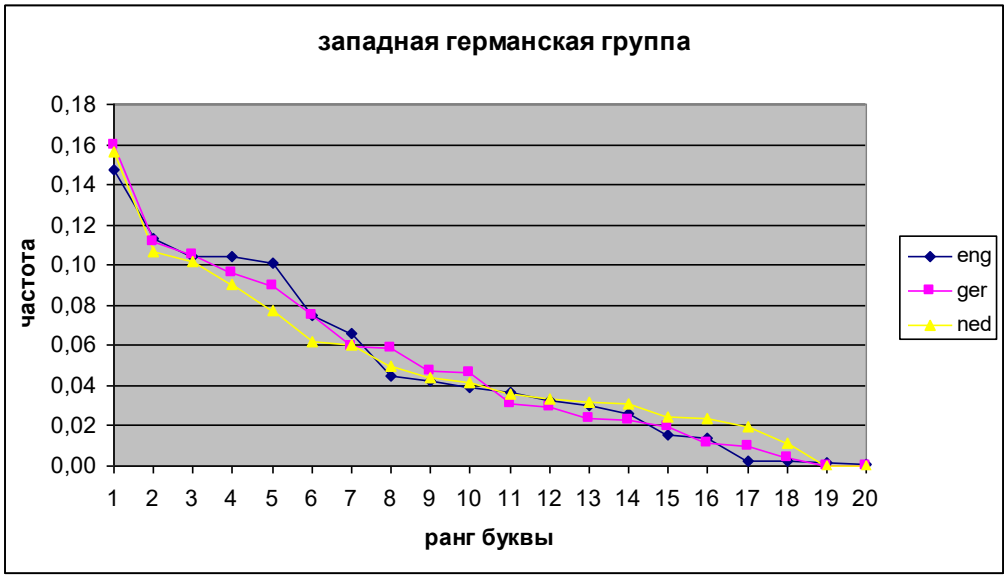

Рис. 1. Распределение частот западногерманской подгруппы

Даже в пределах одной языковой подгруппы на одинаковых местах стоят в общем случае разные буквы. Наиболее часто встречаемые символы (по используемому корпусу) в западногерманской подгруппе приведены в таблице 2.

Таблица 2. Частоты наиболее часто встречаемых согласных букв западногерманской подгруппы

английский	немецкий	голландский
T 0,15	N 0,16	N 0,16
N 0,11	R 0,11	T 0,11
H 0,10	S 0,10	R 0,10
S 0,10	T 0,10	D 0,09

Хотя буквы одинаковых рангов различны, частоты у них примерно равны, т.е. германские языки в пределах своей подгруппы весьма близки, если сравнивать частотно упорядоченные распределения.

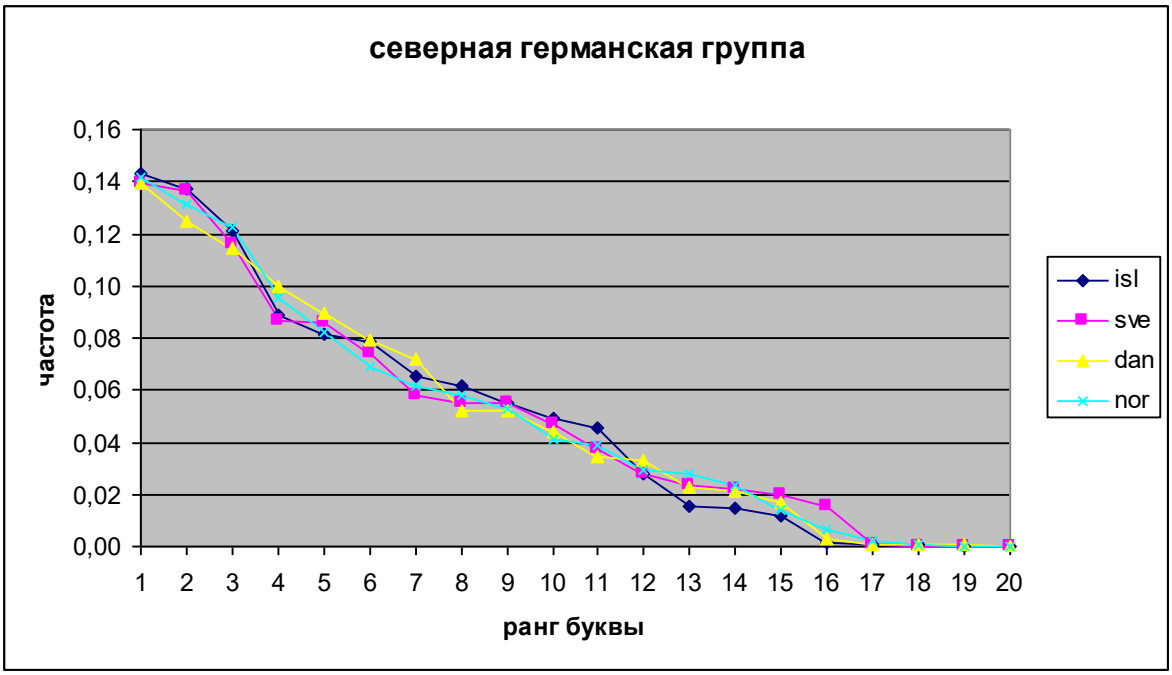

Рис. 2. Распределение частот северной германской подгруппы

Основные отличия германских подгрупп: в северной первая частота заметно ниже, а вторая-четвертая – заметно выше, чем в западной; в северной очень редко употребляются последние четыре буквы, тогда как в западной – только последние две. На рис. 3 эти отличия видны особенно отчетливо. По-видимому, германские подгруппы не объединяются в одну группу. Далее при расчете расстояний между распределениями частот мы покажем это численно.

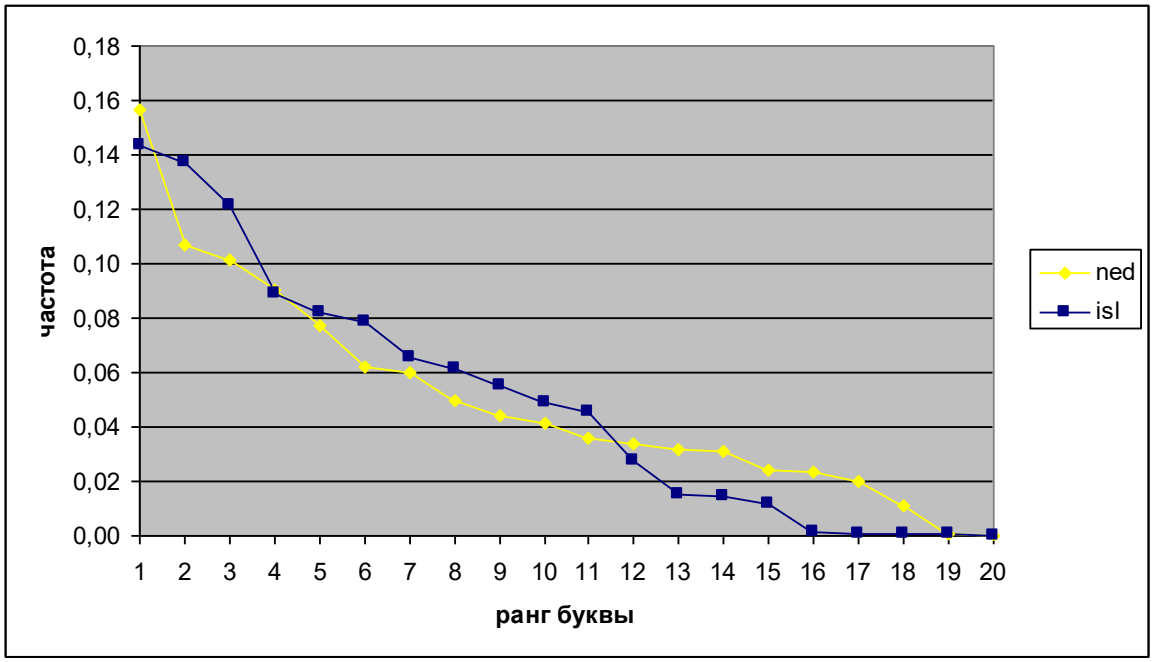

Рис. 3. Сравнение распределений частот
для голландского (западный) и исландского (северный) германских языков

Интересно было бы выяснить, насколько языки балтийской группы (рис. 4) близки к языкам германской группы и к какой именно подгруппе.

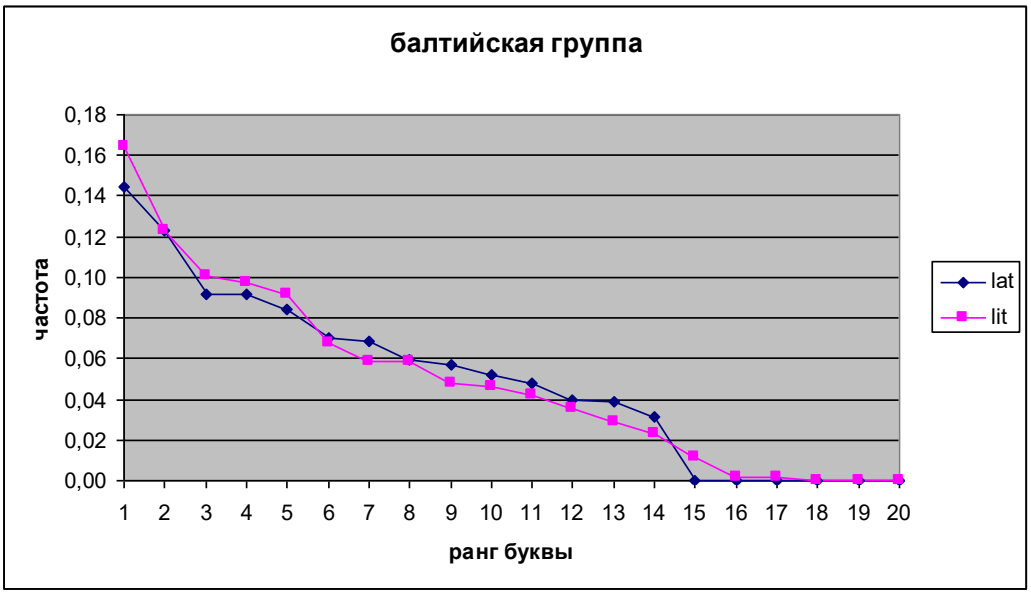

Рис. 4. Распределение частот балтийской группы

Для славянских подгрупп, в отличие от германских, характерно то, что две подгруппы из трех (южная и восточная) объединяются, и только западная подгруппа отделена от остальных. К западной, а не к южной подгруппе оказался ближе хорватский язык, как мы и предположили изначально. Возможно, это связано с различием в алфавитах этих групп. Также, возможно, на различие повлияло то, что группы относятся к разным религиозным конфессиям (православие и католицизм), а письменность первоначально использовалась для записей религиозных текстов.

Также интересно отметить, что один из языков (белорусский) визуально близок не восточным славянам, а западным (рис. 5).

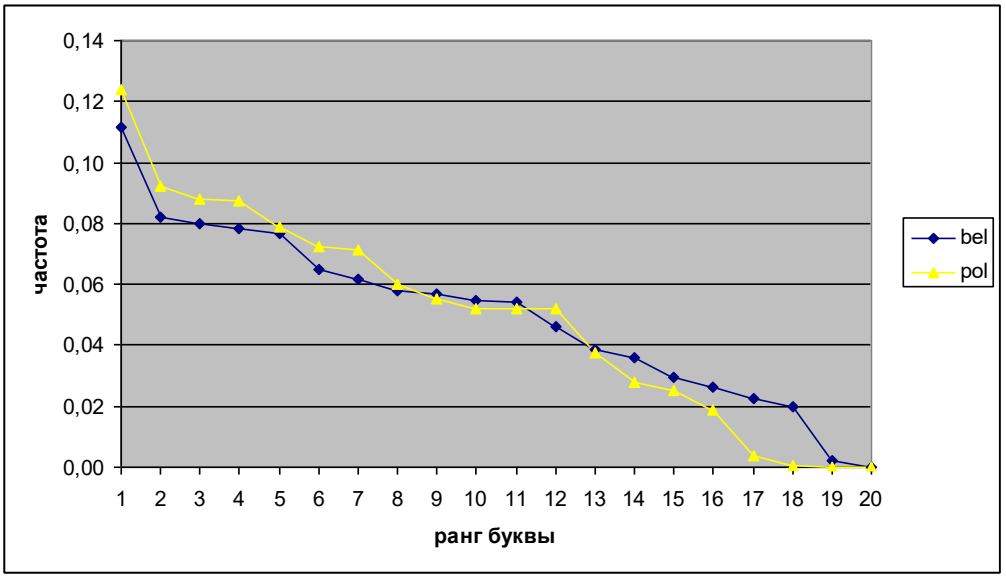

Рис. 5. Сравнение распределений частот
для белорусского (восточный) и польского (западный) славянских языков

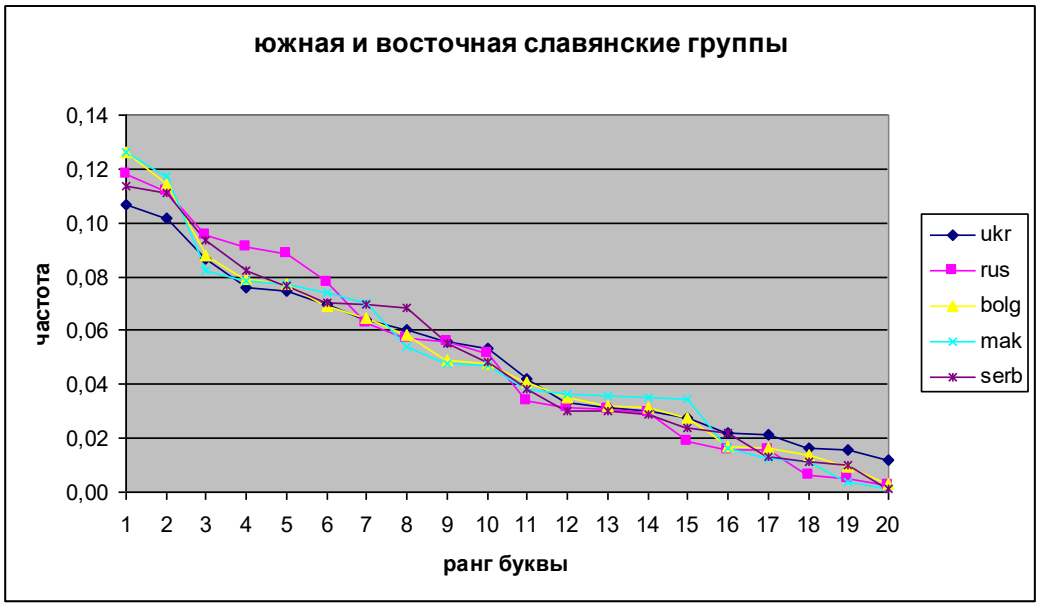

Рис. 6. Распределение частот южной и восточной славянских подгрупп

Отметим, что хотя польский язык технически (в смысле расстояния) близок остальным языкам западнославянской группы, распределение его характерно для языков западногерманской группы (рис. 1). Это показывает, что формальные правила объединения языков в группы неоднозначны. Однако интерес представляет ошибка, которую можно допустить при такой формальной кластеризации.

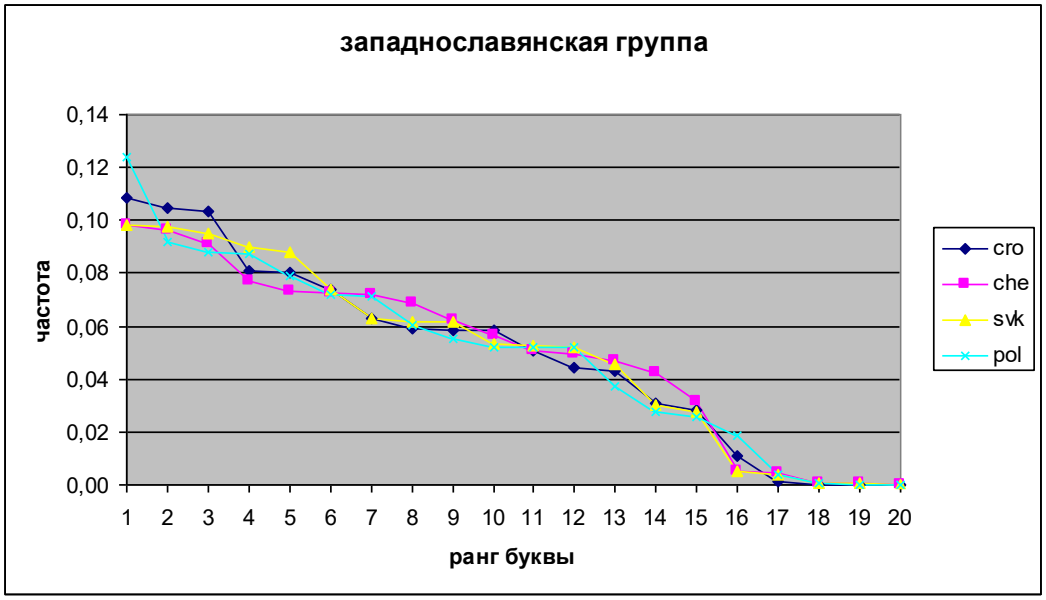

Рис. 7. Распределение частот западнославянской подгруппы

Отличие западнославянской подгруппы от южной и восточной в том, что у них, как и у северной германской подгруппы, почти не используются четыре последних символа – обычно это Q, X, F и W (последний – не в польском). Тогда как у двух других подгрупп даже самые редкие символы используются достаточно заметно.

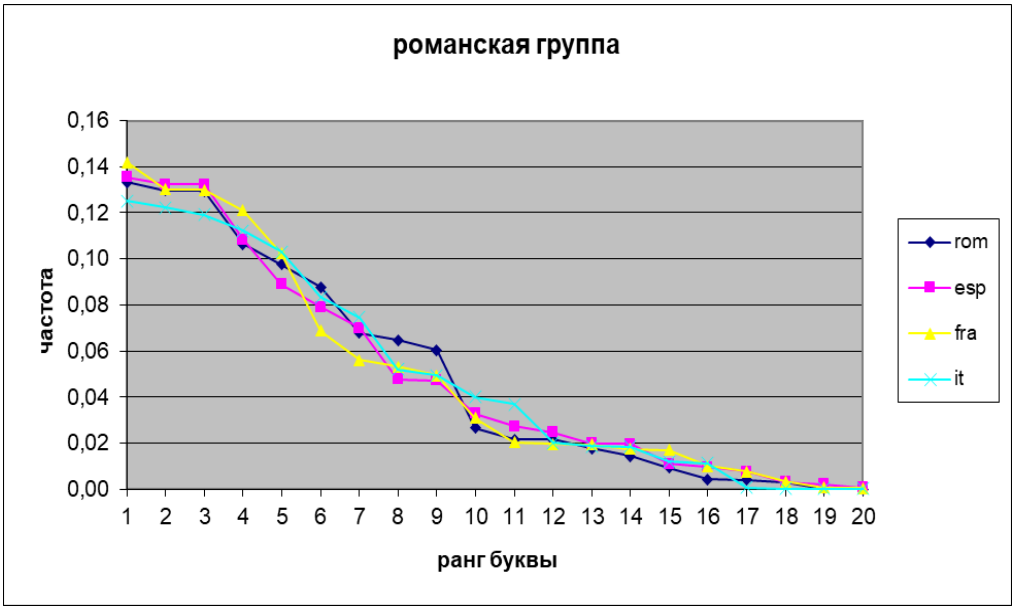

Рис. 8. Распределение частот романской группы

Языки романской группы объединяются почти все, независимо от подгрупп. Однако португальский язык несколько отличается от остальных и оказывается ближе к шведскому (рис. 9), чем к языкам своей подгруппы. Поскольку же выпуклость графиков различна, это свидетельствует о ложной близости таких эталонов.

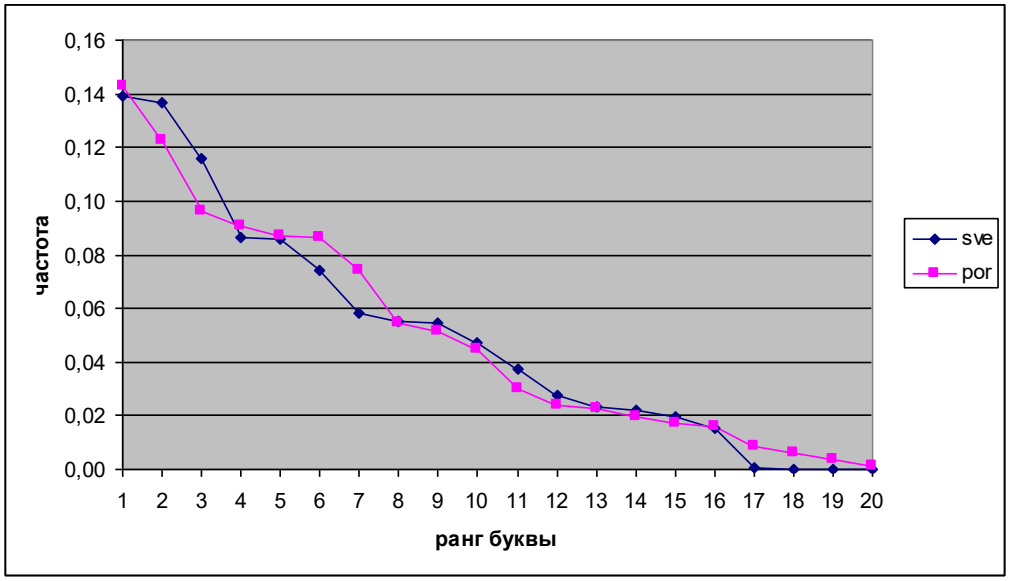

Рис. 9. Сравнение распределений частот
для шведского и португальского языков

Представленные распределения образуют эталонные распределения частот $f(i)$. Теперь надо провести анализ различий между ними.

3. Результаты кластеризации эталонов

Расстояния между эталонными распределениями упорядоченных частот представлены в таблице 3. Матрица корреляций между ними дана в таблице 4.

Пороговые значения ρ^* различны для разных групп. Для западногерманской это значение равно 0,14, для северогерманской – 0,10. Обе подгруппы можно было бы объединить, но этому препятствуют голландский и исландский языки, имеющие заметно большие расстояния до остальных эталонов. Для балтийской группы порог равен 0,11, для южных и восточных славян – 0,14, для западных славян – 0,12, для романской группы – 0,13.

Таблица 3. Расстояния между эталонными распределениями частот европейских языков

	eng	ger	ned	isl	sve	dan	nor	lat	lit	bel	ukr	rus	bolg	mak	serb	cro	che	svk	pol	por	rom	esp	fra	it
eng	0	0,10	0,14	0,18	0,14	0,11	0,13	0,17	0,12	0,31	0,26	0,15	0,19	0,19	0,20	0,21	0,30	0,24	0,21	0,14	0,20	0,15	0,17	0,13
ger		0	0,12	0,16	0,11	0,11	0,12	0,18	0,09	0,29	0,24	0,12	0,17	0,18	0,18	0,20	0,29	0,22	0,21	0,11	0,20	0,16	0,17	0,17
ned			0	0,23	0,18	0,19	0,17	0,20	0,15	0,21	0,18	0,15	0,13	0,15	0,15	0,20	0,28	0,24	0,20	0,17	0,30	0,24	0,25	0,25
isl				0	0,08	0,10	0,08	0,14	0,14	0,33	0,28	0,19	0,22	0,24	0,22	0,21	0,28	0,24	0,22	0,15	0,14	0,14	0,19	0,15
sve					0	0,08	0,07	0,16	0,13	0,29	0,25	0,14	0,18	0,18	0,18	0,18	0,28	0,22	0,19	0,11	0,18	0,14	0,16	0,14
dan						0	0,08	0,14	0,11	0,33	0,28	0,15	0,20	0,19	0,21	0,21	0,28	0,23	0,21	0,09	0,14	0,10	0,15	0,10
nor							0	0,13	0,10	0,30	0,26	0,16	0,18	0,20	0,19	0,19	0,28	0,22	0,21	0,14	0,15	0,12	0,14	0,14
lat								0	0,11	0,24	0,23	0,16	0,17	0,18	0,19	0,15	0,20	0,16	0,14	0,16	0,24	0,23	0,26	0,22
lit									0	0,30	0,27	0,17	0,19	0,21	0,22	0,20	0,28	0,22	0,21	0,15	0,21	0,19	0,20	0,18
bel										0	0,12	0,19	0,15	0,16	0,17	0,15	0,16	0,17	0,13	0,29	0,41	0,39	0,40	0,37
ukr											0	0,14	0,09	0,14	0,09	0,15	0,18	0,18	0,16	0,24	0,36	0,34	0,36	0,32
rus												0	0,10	0,13	0,09	0,15	0,22	0,15	0,15	0,11	0,24	0,20	0,24	0,20
bolg													0	0,06	0,08	0,15	0,20	0,19	0,14	0,17	0,30	0,26	0,28	0,24
mak														0	0,10	0,16	0,19	0,20	0,14	0,16	0,31	0,24	0,29	0,23
serb															0	0,15	0,18	0,18	0,14	0,17	0,28	0,26	0,30	0,25
cro																0	0,10	0,08	0,10	0,23	0,30	0,29	0,32	0,25
che																	0	0,08	0,12	0,29	0,36	0,36	0,40	0,33
svk																		0	0,10	0,24	0,30	0,31	0,34	0,28
pol																			0	0,20	0,31	0,29	0,33	0,25
por																				0	0,17	0,13	0,17	0,14
rom																					0	0,09	0,11	0,12
esp																						0	0,09	0,11
fra																							0	0,13
it																								0

Желтым цветом выделены полносвязные диагональные блоки языковых подгрупп. Фиолетовым цветом отмечены пары языков, которые оказались близки между собой, но относятся к разным подгруппам.

Таблица 4. Коэффициенты корреляции между эталонными распределениями частот европейских языков, $c*10^3$

	ger	ned	isl	sve	dan	nor	lat	lit	bel	ukr	rus	bolg	mak	serb	cro	che	svk	pol	por	rom	esp	fra	it
eng	988	979	968	976	988	980	971	987	953	972	982	979	972	966	950	909	937	963	983	969	978	978	982
ger		990	976	980	985	981	970	992	956	979	980	986	974	973	945	906	922	960	986	966	970	973	966
ned			959	968	969	973	958	984	952	960	959	978	968	956	927	885	896	948	970	943	959	959	943
isl				994	991	993	973	978	941	992	981	986	970	985	966	929	940	954	982	977	980	965	974
sve					990	995	970	980	942	990	985	989	978	985	966	924	937	952	984	977	982	974	975
dan						993	977	983	951	990	991	988	979	985	970	937	953	967	990	985	990	980	989
nor							976	984	944	986	983	987	972	984	968	931	944	956	979	982	989	981	978
lat								987	963	979	976	982	974	973	967	949	954	974	970	945	949	937	953
lit									959	978	976	988	977	971	953	919	934	967	979	956	962	962	959
bel										960	959	962	957	957	965	951	958	984	945	916	917	912	932
ukr											993	992	983	996	973	946	952	967	990	976	973	962	977
rus												986	978	989	971	942	959	964	989	981	979	974	986
bolg													995	990	964	937	940	966	990	964	969	960	966
mak														981	961	941	939	965	984	949	955	943	954
serb															969	948	950	965	985	976	971	961	975
cro																985	990	980	948	943	943	928	954
che																	985	968	913	906	898	878	915
svk																		974	924	930	922	913	944
pol																			956	930	931	920	946
por																				977	977	967	978
rom																					991	987	988
esp																						990	991
fra																							984

Из таблицы 3 следует, что почти все языки, входящие в определенную языковую подгруппу, находятся между собой на расстоянии не большем, чем $\rho \le \rho^* = 0{,}14$. Таким способом выделены шесть языковых подгрупп из восьми, две подгруппы оказались близки для славянской группы (южная и восточная) и две для романской группы.

Два языка не дали полного соответствия элементам своей группы.

1. Белорусский язык из своей восточнославянской группы оказался близок только украинскому $\rho = 0{,}12$. Из всей остальной группы языков он близок польскому языку с расстоянием $\rho = 0{,}13$.

2. Португальский язык из языков романской группы близок испанскому $\rho = 0{,}13$ и итальянскому $\rho = 0{,}14$. Однако ближайшими соседями португальского языка оказались немецкий и шведский языки с расстоянием $\rho = 0{,}11$.

Также можно заметить, что близкими оказались некоторые языки из разных групп. Например, литовский и итальянский языки близки английскому и датскому языкам, латышский близок норвежскому. Всего в таблице 2 указана 41 пара близких эталонов из 276 пар, которые, однако, относятся к разным языковым группам, то есть ошибка ложного принятия в класс составляет примерно 15%. Под близостью понимается расстояние, не превосходящее выбранного уровня 0,14. Это показывает, что формальное объединение языков в одну группу по близости распределений упорядоченных частот согласных букв неоднозначно. Для правильной кластеризации требуются некоторые дополнительные соображения – исторические, лингвистические и т.п.

Тем не менее важным результатом проведенной работы явилось то, что за исключением двух языков из 24 подтвердилась лингвистическая гипотеза о том, что близкие языки индоевропейской семьи имеют сходный согласный остов своих лексиконов, а различия связаны главным образом с использованием гласных, т.е. с огласовкой. Тем самым более 90% эталонов удовлетворяют гипотезе близости распределений, а ошибка первого рода (пропуск цели) составила 8%.

Для сравнения с методом анализа близости эталонов в норме L1 полезно рассмотреть матрицу корреляций между распределениями (таблица 4). С одной стороны, близость эталонов в смысле ближайшего соседа еще не означает высокой корреляции между эталонами, поэтому результаты кластеризации двумя методами не обязаны совпадать. С другой стороны, найденные распределения упорядоченных частот в одной группе имеют примерно одинаковые графики, поэтому можно предположить, что анализа корреляций будет вполне достаточно для нахождения языковых групп.

Однако оказалось, что корреляционный анализ приводит к менее точным результатам по сравнению с ближайшими соседями. Действительно, если предположить, что в одну группу объединяются эталоны с корреляцией между собой не менее 0,990 (число 990 в таблице 4), то правильно сгруппировать получается только 12 языков из 24. Если же понизить уровень кластеризации до 0,978, то 21 эталон попадает в свою языковую группу, но зато ошибка ложного

принятия (близость между чужими эталонами) на этом уровне составила 71 пару из 276, т.е. 26%. Следовательно, анализ распределений в гистограммной норме оказывается существенно более точным.

Полученные результаты дополняют данные о близости европейских языков, полученные ранее в [8].

4. Графы ближних и дальних соседей

Интересно попробовать провести кластеризацию эталонов методом графа ближайшего соседа в соответствии с расстояниями между эталонами. Правило объединения эталонов в группу следующее. Берем первый эталон (eng) и ищем ближайший к нему. Из таблицы 3 следует, что это эталон (ger). Далее ищем ближайшего соседа для (ger). Им оказывается эталон (lit). И так далее. Кластером будем называть каждый связный граф.

Выяснилось (рис. 10), что таким путем получаются пять групп, в которые входят следующие языки:

1. английский, немецкий, голландский, литовский, латышский;
2. исландский, шведский, норвежский, датский, португальский;
3. белорусский, украинский, болгарский, македонский, сербский, русский;
4. чешский, словацкий, хорватский, польский;
5. румынский, испанский, французский, итальянский.

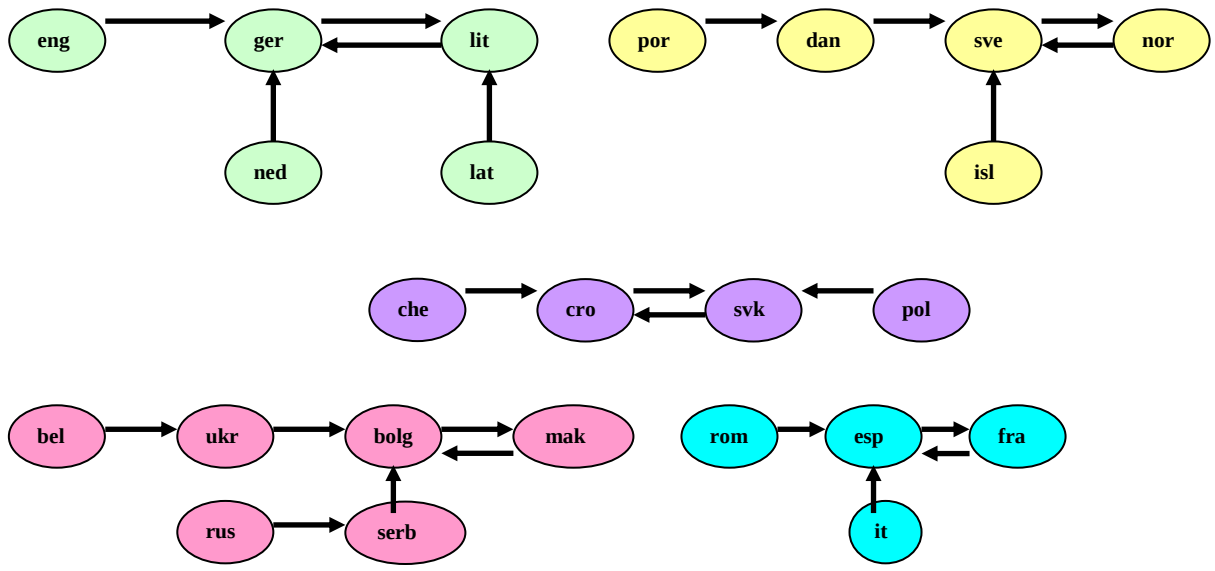

Рис. 10. Граф ближайших соседей эталонов частот

Таким образом, метод ближайшего соседа дает примерно те же результаты, что и парная кластеризация, однако есть некоторые отличия. Три языка (литовский, латышский и португальский) включены не в свои подгруппы. Это несколько хуже, чем объединение методом попарной близости всех эталонов, но тем не менее результат достаточно точный.

При этом центрами групп (это вершины с максимальной степенью по входящему ребру) являются языки: немецкий, шведский, болгарский, хорватский, испанский.

Отметим, что в работах [9, 10] были представлены численные оценки вероятности того, что граф ближайших соседей, построенный по случайной матрице расстояний, имеет определенную структуру: число несвязных фрагментов, распределение числа вершин по фрагментам, распределение вершин по степеням. Отклонение от расчетных показателей для случайных графов трактуется как наличие связи между соседями. В частности, для случайного графа ближайших соседей из N вершин наиболее вероятное число несвязных фрагментов равно N/4. Для N = 24 это число 6. Естественно, что в данной задаче расстояния между эталонами не случайны. Однако на рис. 10 показано 5 фрагментов, что в целом близко к случайности. Важным аспектом использования графов для оценки неслучайности соседей является то, что для случайных данных статистики графов ближних и дальних соседей совпадают.

Рассмотрим граф дальних соседей для матрицы расстояний из таблицы 3.

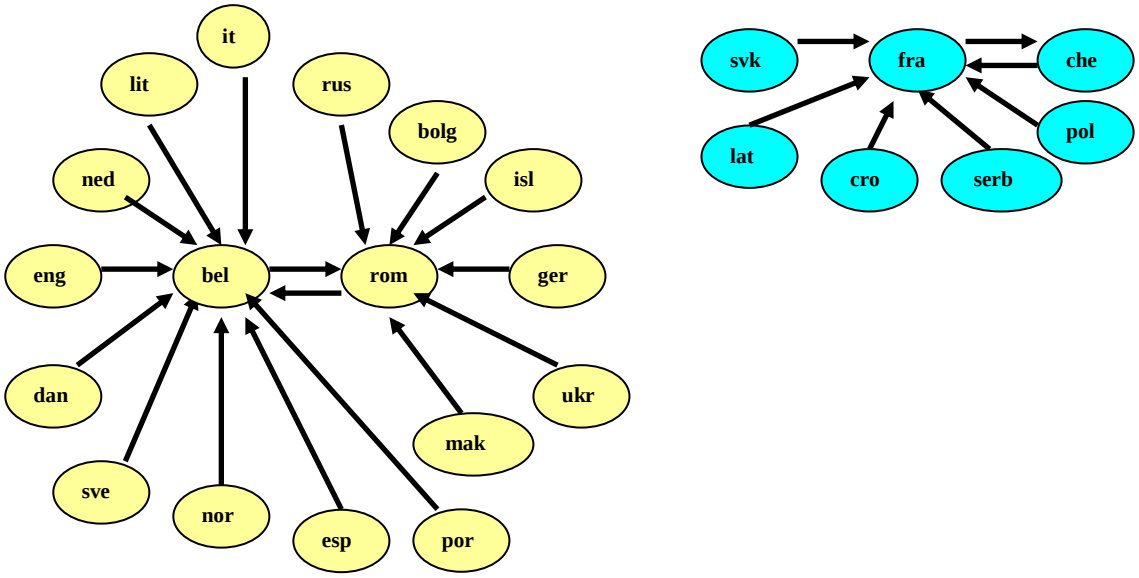

Рис. 11. Граф дальних соседей эталонов частот

Этот граф состоит из двух фрагментов и существенно отличается от графа на рис. 10. Следовательно, как и должно быть, гипотеза о том, что языковые эталоны независимы, отклоняется. По оценкам работы [9] вероятность независимости данных меньше, чем 0,001. Языки, наиболее далеко отстоящие от остальных в смысле статистики символов, – это белорусский, румынский и французский.

Этот пример приведен для того, чтобы показать важность наличия независимых критериев проверки связи между набором числовых величин. Вероятность независимости оценивается по наихудшему результату, что позволяет избежать ошибки ложного принятия.

Определенный интерес вызывает также вопрос о связности графа в целом, если соединять между собой вершины, находящиеся на расстоянии, не большем заданного значения (порога кластеризации).

Минимальное расстояние в таблице 3 равно 0,06 и отвечает паре языков «болгарский–македонский». Поэтому если выбрать порог меньше этого значения, то граф будет представлять 24 изолированные вершины. Начиная с порога 0,06, появляется первое ребро. При пороге 0,07 появляется второе ребро между языками «шведский–норвежский». При пороге 0,08 имеем следующую картинку (рис. 12):

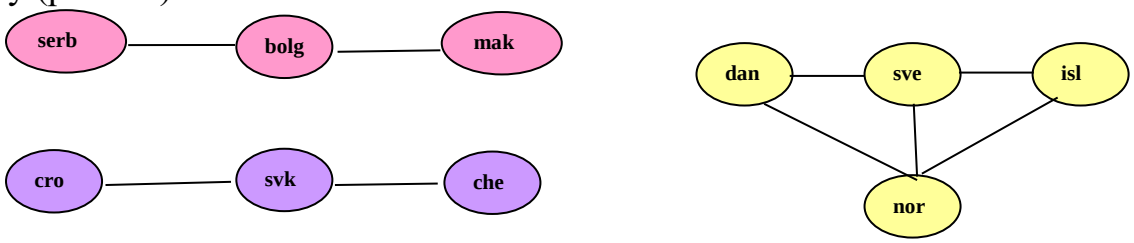

Рис. 12. Фрагменты графа кластеризации на уровне порога 0,08

Остальные вершины остались изолированными.

Продолжая этот процесс дальше, т.е. увеличивая порог объединения, получаем граф, содержащий все меньшее число изолированных вершин. На уровне порога 0,12 граф состоит из двух фрагментов: западнославянского (Хорватия, Чехия, Словакия, Польша), который при этом оказывается полносвязным, и остальных вершин, образующих связный подграф, но не полносвязный. В этом втором фрагменте есть полносвязный подграф из стран северогерманской группы (Исландия, Дания, Швеция, Норвегия).

При пороге 0,13 граф языковых эталонов становится связным, так как появляется связь между белорусским и польским языками. Также образуется еще один полносвязный подграф романской группы (кроме, как уже говорилось, Португалии).

На уровне кластеризации 0,14 все языковые группы образуют полносвязные подграфы.

Заключение

В данной работе были рассмотрены два метода кластеризации данных, использующих расстояние между элементами множества: ближайшие пары и ближайшие соседи. Метод ближайших соседей оказался чуть менее точным в формировании групп близкородственных языков. Это связано с тем, что граф ближайших соседей не может быть полносвязным (если число вершин больше двух), так как ближайшая точка единственна. Лингвистические же группы представляют собой объединение эталонов, которые попарно близки между собой на уровне данного порога. Отыскание такого порога представляет собой задачу, решаемую численно для каждой заданной группы объектов. В данном

примере языков индоевропейской семьи порог объединения оказался равным 0,14.

Также работа имела целью продемонстрировать результаты применения бенчмарка статистик графов ближайших соседей «в обратную сторону»: взять заведомо связанные между собой элементы множества и показать, что бенчмарк действительно дает малую вероятность их независимости.

Разделы 2 и 3 написаны М.Ю. Кислицыной, остальные разделы написаны совместно.

Благодарности

Авторы считают своим приятным долгом поблагодарить коллектив лицеистов 8 класса АНОО «Физтех-лицей» им. П.Л. Капицы: Белякову Маргариту, Босову Александру, Смелова Даниила и Федянова Владислава, а также куратора их научной работы, преподавателя математики Дмитрия Сергеевича Герасимова – за поднятие этой темы и большую техническую помощь в подборе корпуса текстов и их предобработке.

Литература

1. Рассел С., Норвиг П. Искусственный интеллект. Современный подход. – М.: Вильямс. 2007. – 1480 с.

2. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. – М.: Вильямс, 2011. – 528 с.

3. Hovy E., Lavid Ju. Towards a science of corpus annotation // International Journal of Translation, 2010. V. 22. No 1. P. 1-25.

4. Арутюнов А.А., Борисов Л.А., Зенюк Д.А., Ивченко А.Ю., Кирина-Лилинская Е.П., Орлов Ю.Н., Осминин К.П., Федоров С.Л., Шилин С.А. Статистические закономерности европейских языков и анализ рукописи Войнича // Препринты ИПМ им. М.В. Келдыша, 2016. № 52, 36 с.

5. Яглом А.М., Яглом И.М. Вероятность и информация. – М.: КомКнига, 2007. – 512 с.

6. Лиланд Р. Криптологическая математика. – Математическая ассоциация Америки, 2000. – 199 с.

7. Лингвистический энциклопедический словарь. – М.: Советская энциклопедия, 1990. – 683 с.

8. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.

9. Кислицын А.А., Орлов Ю.Н. Исследование статистик графов ближайших соседей // Препринты ИПМ им. М.В. Келдыша, 2021. № 85, 23 с.

10. Кислицын А.А. Моделирование графов ближайших соседей для оценки вероятности независимости выборочных данных // Математическое моделирование. 2023. Т. 35. № 7. С. 63-82.