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

Препринты ИПМ • Препринт № 32 за 2017 г.

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

Орлов Ю.Н., Шилин С.А.

Статистическое распознавание языка текста по частоте буквосочетаний

Рекомендуемая форма библиографической ссылки: Орлов Ю.Н., Шилин С.А. Статистическое распознавание языка текста по частоте буквосочетаний // Препринты ИПМ им. М.В.Келдыша. 2017. № 32. 21 с. doi:10.20948/prepr-2017-32
URL: http://library.keldysh.ru/preprint.asp?id=2017-32

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

Ю.Н. Орлов, С.А. Шилин

Статистическое распознавание языка текста по частоте буквосочетаний

Москва — 2017

Орлов Ю.Н., Шилин С.А.

Статистическое распознавание языка текста по частоте буквосочетаний

Исследуются статистические свойства текстов, написанных на языках индоевропейской семьи, с целью разработки индикаторов распознавания языка или языковой группы. В качестве индикаторов рассматриваются: показатель Херста для некоторой специальной статистики, показывающей однородность звучания текста; частоты наиболее употребляемых буквосочетаний в текстах на разных европейских языках; эталонные распределения *n*-грамм. Точность бинарного распознавания была достигнута на уровне 0,99.

Ключевые слова: частоты буквосочетаний, распознавание языка текста

Orlov Yu.N., Shilin S.A.

Statistical text language recognition with the use of n-gram frequency

Statistical properties of European language texts are investigated with the use of recognition procedure for n-gram distribution patterns. The numerical algorithm is constructed for analysis Hurst exponent for letter distance distributions of the text fragment. The accuracy of binary recognition is estimated as 0,99.

Key words: text language recognition, n-gram frequency

Работа выполнена при поддержке гранта РФФИ, проект № 16-01-00342

Содержание

Введение

В настоящем исследовании представлены результаты по статистической классификации языков индоевропейской семьи. Задача распознавания языка текста была поставлена [1], где были определены некоторые базовые статистики, такие, как распределения упорядоченных буквосочетаний. Продолжением этой работы явилась статья [2], в которой исследовался некий зашифрованный текст, так называемый манускрипт Войнича, на предмет соответствия какому-либо из европейских языков. Было сформулировано достаточно обоснованное предположение, что манускрипт мог быть написан на двух и более языках. В продолжение исследований о распределении однобуквенных и двухбуквенных сочетаний в европейских языках и изучении похожих закономерностей в манускрипте мы рассмотрим 3-х и 4-х символьные последовательности на предмет того, насколько хорошо они могут идентифицировать тот или иной язык.

Рассматриваются следующие языки: болгарский, хорватский, чешский, сербский, польский, датский, голландский, шведский, норвежский, английский, немецкий, испанский, румынский, итальянский, французский.

Также мы проведем тестирование гипотезы о том, что так называемый показатель Херста, вернее, его выборочное распределение, может служить индикатором однородности языкового состава текста. Статистические эксперименты были поставлены на двух достаточно представительных корпусах текстов: собрании сочинений классической литературы на каждом из вышеперечисленных пятнадцати языков, а также статей Википедии на этих языках. Оба корпуса дали совпадающие результаты.

Для проведения анализа текста на одном из языков была проведена компоновка статей Википедии для соответствующего языка, средняя длина такого объединенного текста составила 213 миллионов символов.

По аналогии с манускриптом Войнича мы будем рассматривать тексты без огласовки, написанные одним алфавитом (латиница). Например, имеем исходный текст на сербском языке:

“Папа Јован VIII

Папа Јован VIII (латински: Ioannes VIII; умро 16. децембра 882) је био 107. папа од 13. децембра 872. године до своје смрти. Сматра се за једног од најспособнијих папа у 9. веку. Велики део свог понтификата Јован је провео у борби са муслиманским освајачима који су из својих упоришта у јужној Италији организовали походе ка северу угрожавајући тако економију папске баштине.”…

Результат первоначальной обработки имеет вид

ппјвнппјвнлтнскмрдцмбрјбппддцмбргдндсвјсмртсмтрсзјднгднјспсбнјхппвквлкдсвгпнтфктјвнјпрвбрбсмслмнскмс…,

а после перевода в латинскую транслитерацию получаем:

ppjvnppjvnltnskmrdcmbrjbppddcmbrgdndsvjsmrtsmtrszjdngdnjspsbnjhppvkvlkdsvgpntfktjvnjprvbrbsmslmnskms.

Предыдущие исследования показали, что, анализируя только однобуквенные распределения, достаточно 30000 символов, чтобы с точностью около 95% определять язык текста. В этой работе мы разобьем исследуемые тексты на фрагменты по N символов. Далее выяснится, какая длина

содержащихся в этих фрагментах n-грамм является оптимальной для хорошего качества предсказания. В настоящей работе исследуется эффективность обучения машинного классификатора на векторах трех и четырехбуквенных последовательностей этих фрагментов. Оказалось, что предлагаемый способ распознавания может дать возможность определять язык даже для малых фрагментов текста (до 100 символов) с большой точностью, поскольку удается выделить некоторые n-граммы, которые с высокой достоверностью специфичны для того или иного языка.

Непосредственно для классификации n-грамм будем использовать логистическую регрессию, которая очень хорошо работает на бинарных данных, а сам классификатор будем обучать по принципу One-Vs-Rest (один против всех), этот способ включает в себя обучение классификатора для каждого класса с образцами этого класса в качестве положительных объектов (индикатор «1») и всех других образцов как отрицательных (индикатор «0»).

1. Частотное распределение букв в европейских языках

Результаты обработки большого корпуса текстов позволили получить стационарные паттерны распределения символов в текстах на европейских языках, записанных без огласовки. Эти распределения представлены ниже на диаграммах.

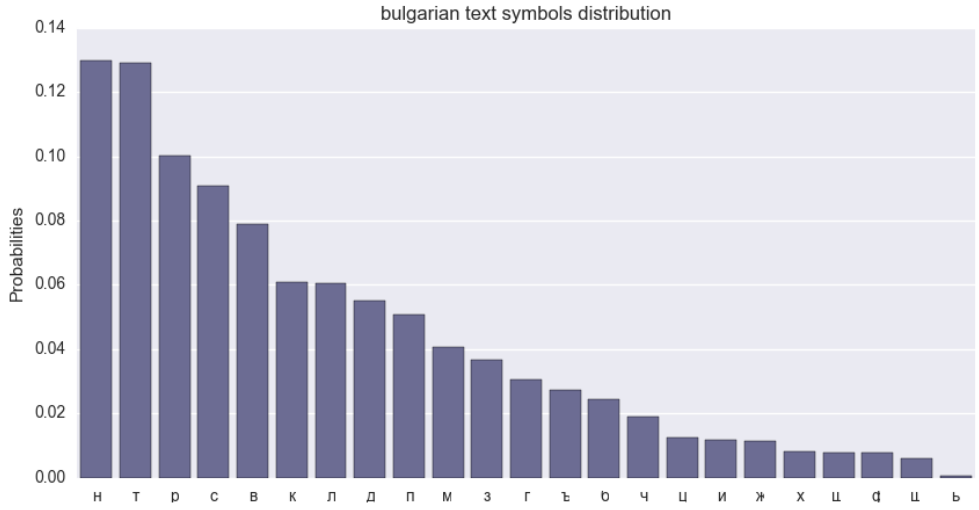
Важно понимать, что некоторый конкретный текст, написанный на одном из рассмотренных языков, может весьма заметно отличаться от своего паттерна, т.е. собственно по значению частоты встречаемости символа в тексте нельзя с уверенностью сказать, какой это язык. Однако с высокой вероятностью буквенное частотно-упорядоченное распределение для исследуемого текста оказывается в определенной норме ближе именно к своему эталону, а не к эталону чужого языка.

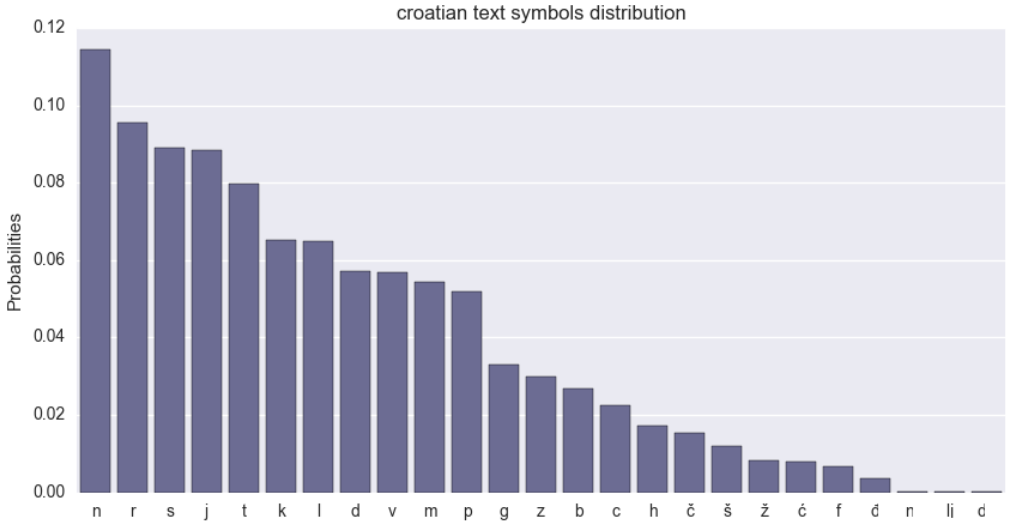
Логарифмическая аппроксимация распределения частот [2] для построенных паттернов варьируется на уровне 0,95-0,98. Существуют определенные групповые отличия, которые и позволяют достаточно хорошо определять язык текста, даже если он записан шифром прямого соответствия.

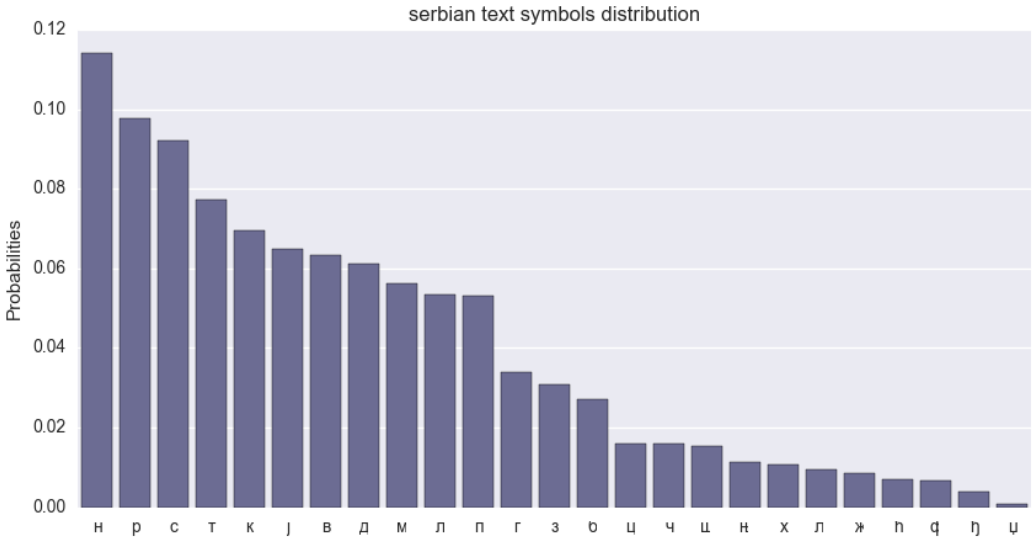
Так, языки романской группы имеют близкий профиль первых пяти букв (это медиана распределения), позволяющий отделить эту группу от остальных, и различающийся «хвост», по которому уже можно распознать язык внутри самой группы.

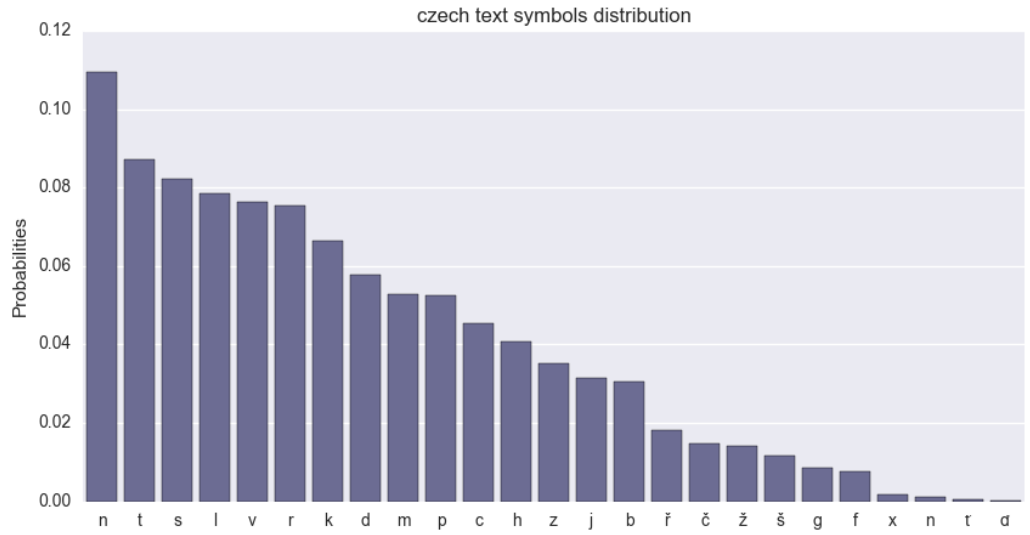
Для языков славянской группы характерным является близкий профиль распределения первой половины букв (примерно квантиль порядка 0,75).

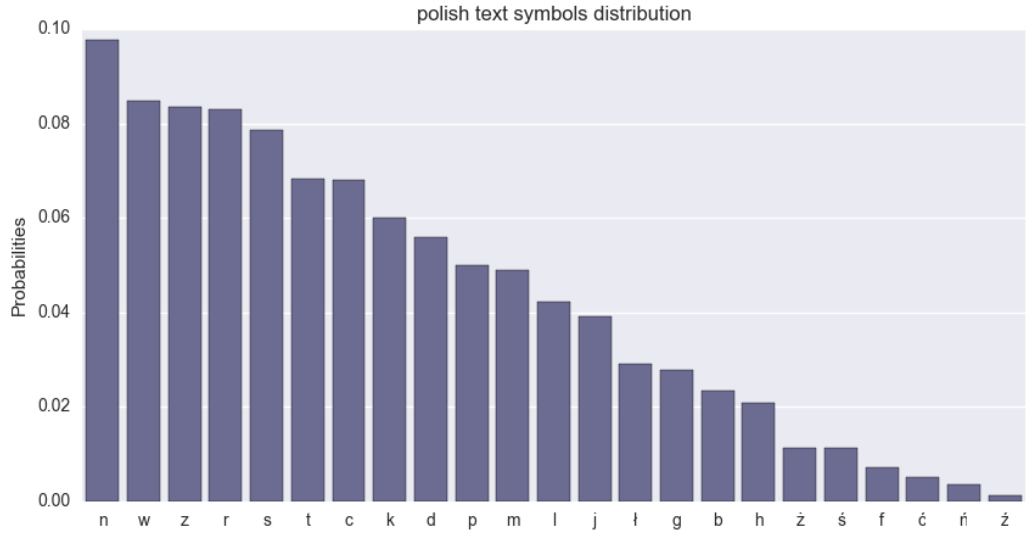
Для языков германской группы характерно пренебрежимо малое использование последних четырех символов (суммарно на уровне менее 0,01), подгруппы же различаются положением медианы и интерквартильным размахом.

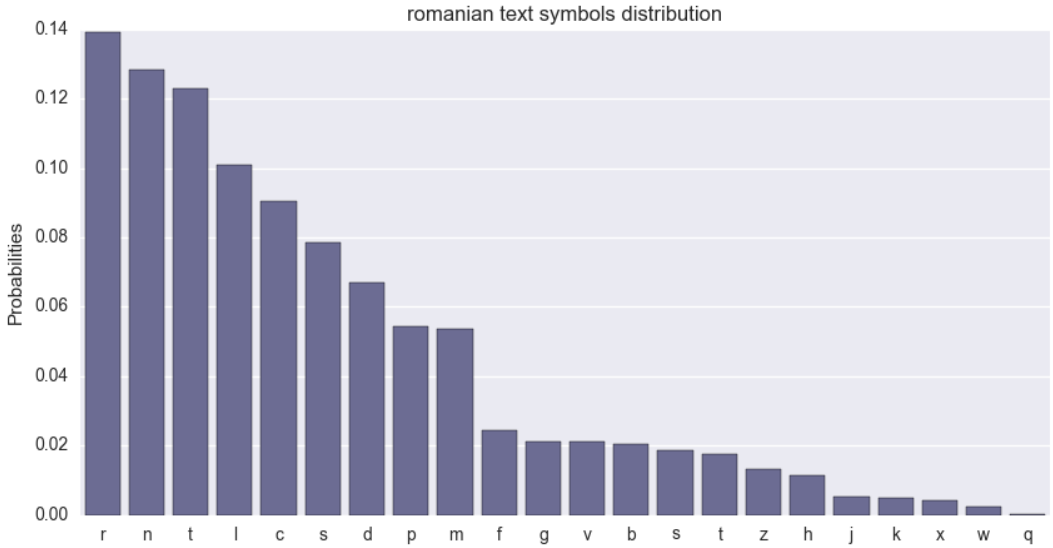

bulgarian text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
н т р с в к л д п м з г ъ ю ч ц и ж х щ ф ш ь

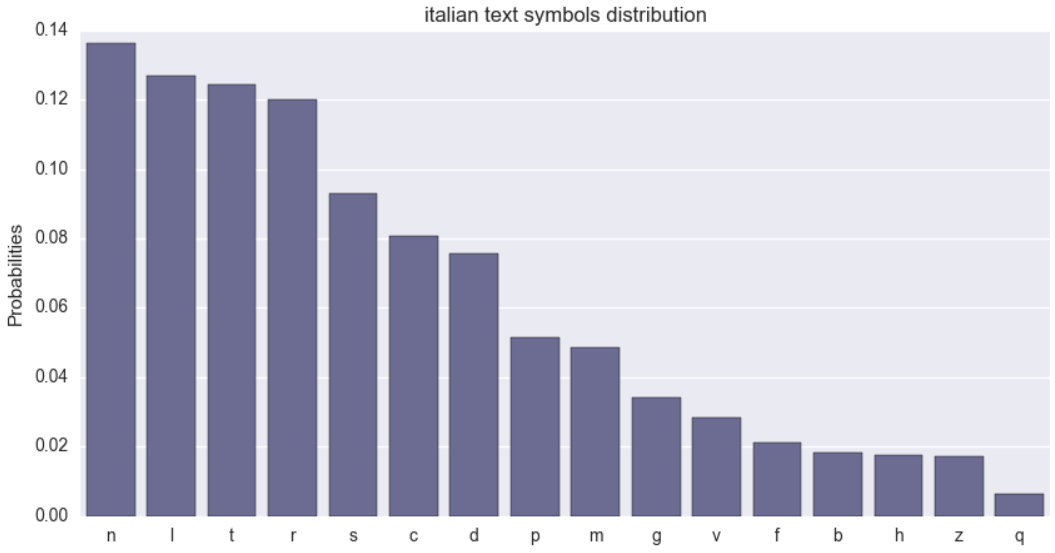

croatian text symbols distribution
Probabilities
0.12
0.10
0.08
0.06
0.04
0.02
0.00
n r s j t k l d v m p g z b c h č š ž ć f đ ŋ lj d

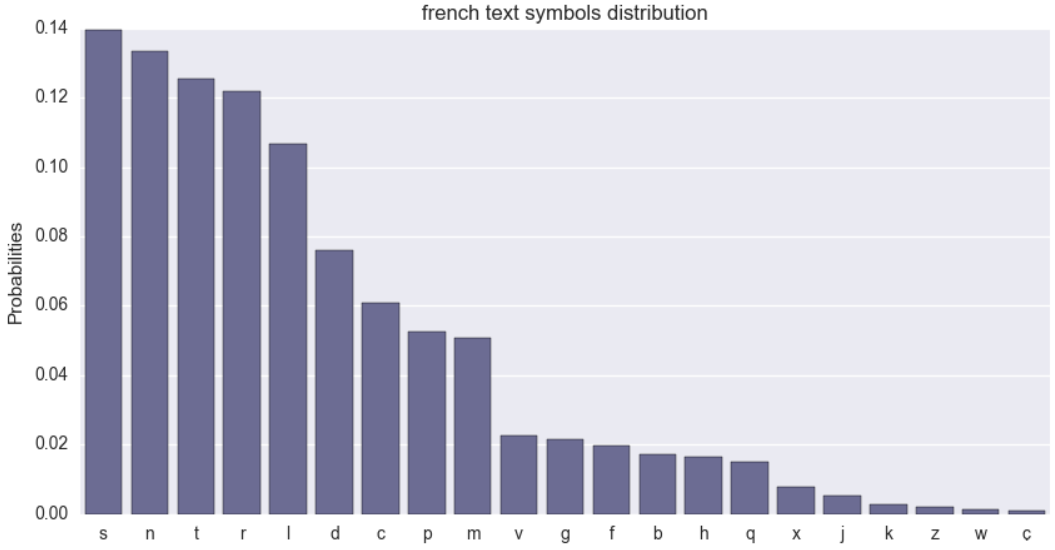

serbian text symbols distribution
Probabilities
0.12
0.10
0.08
0.06
0.04
0.02
0.00
н р с т к ј в д м л п г з ю ц ч ш н х л ж њ ф ђ џ

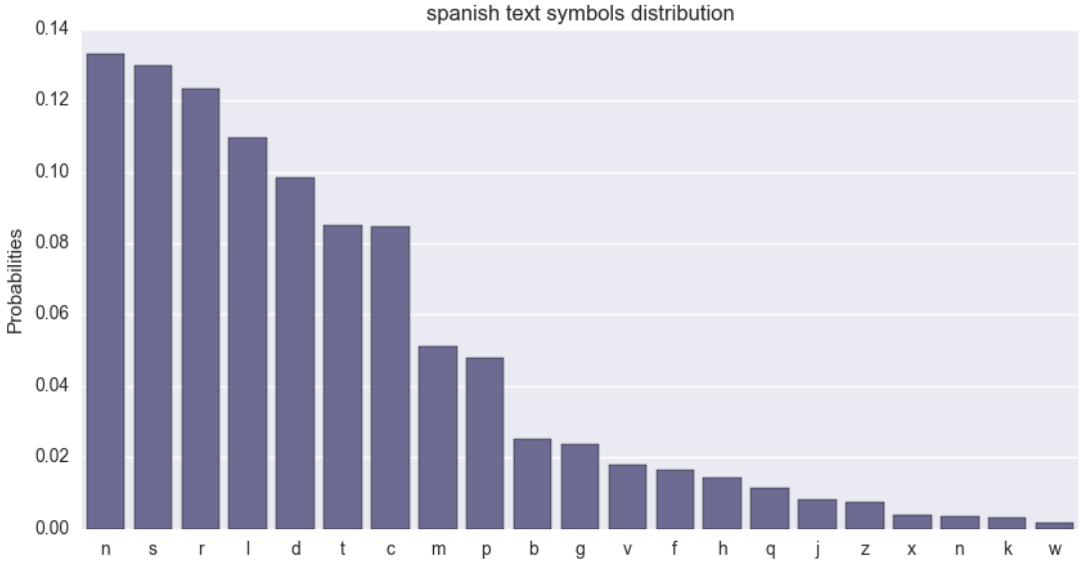

czech text symbols distribution
Probabilities
0.12
0.10
0.08
0.06
0.04
0.02
0.00
n t s l v r k d m p c h z j b ř č ž š g f x n ť ď

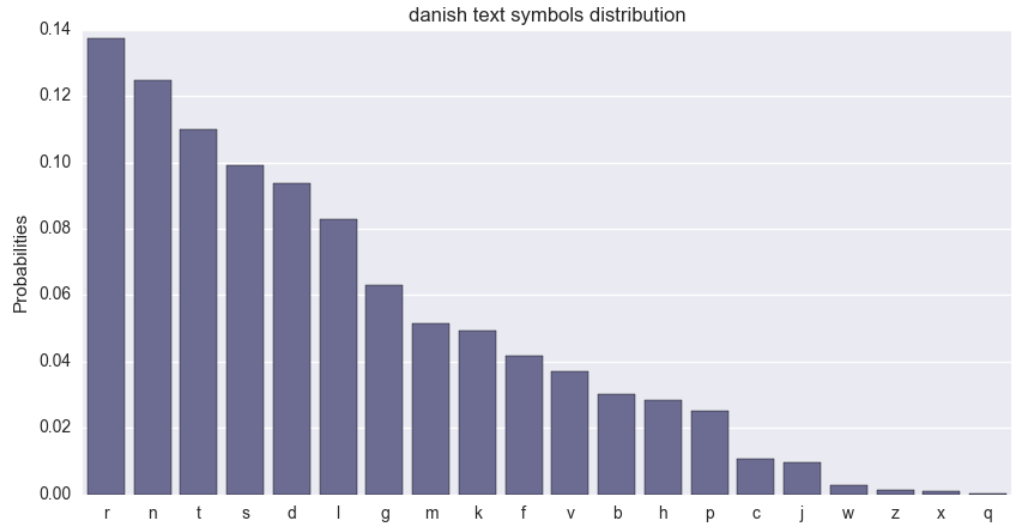

polish text symbols distribution
Probabilities
0.10
0.08
0.06
0.04
0.02
0.00
n w z r s t c k d p m l j ł g b h ż ś f ć ń ż

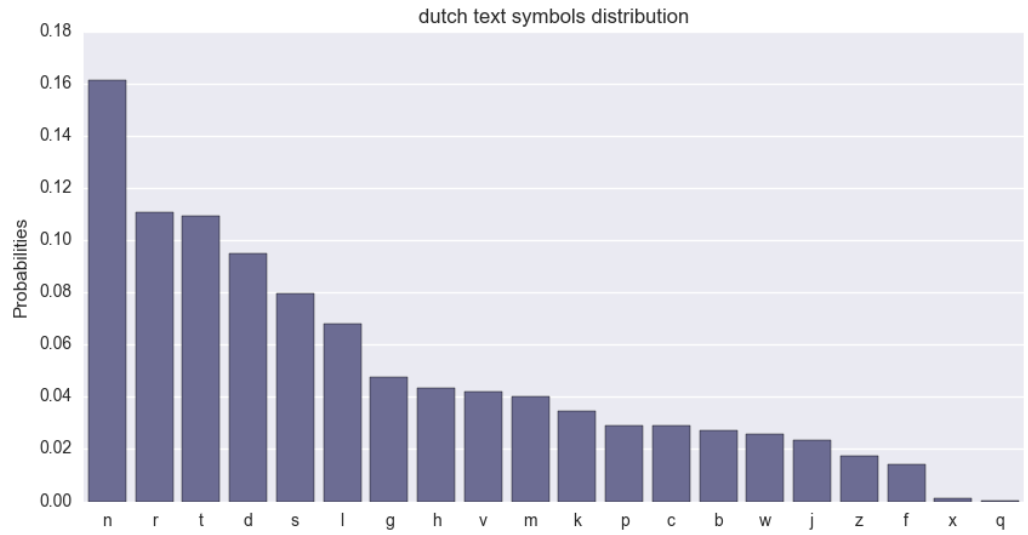

romanian text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
r n t l c s d p m f g v b s t z h j k x w q

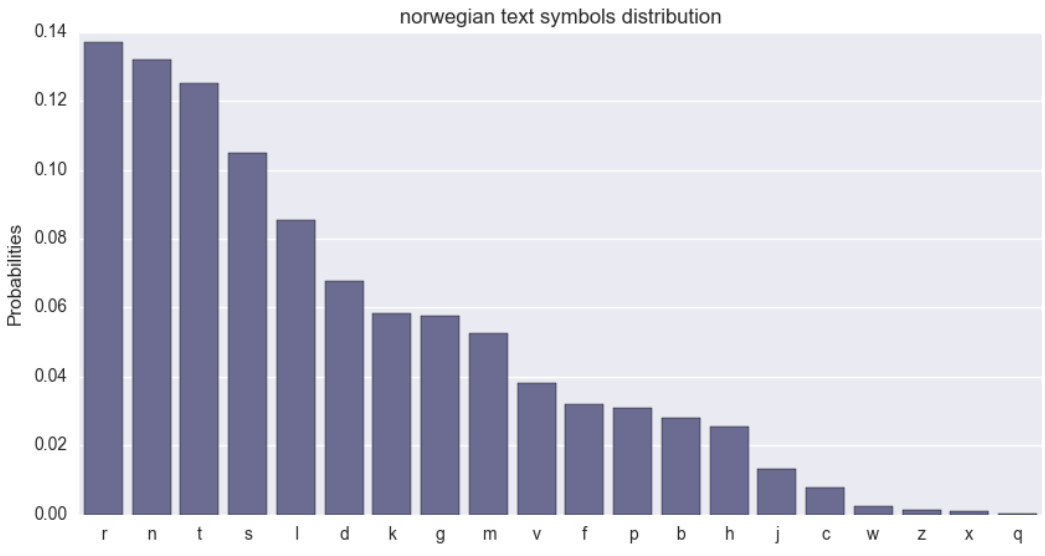

italian text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
n l t r s c d p m g v f b h z q


french text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
s n t r l d c p m v g f b h q x j k z w ç


spanish text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
n s r l d t c m p b g v f h q j z x n k w


danish text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
r n t s d l g m k f v b h p c j w z x q


dutch text symbols distribution
Probabilities
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
n r t d s l g h v m k p c b w j z f x q


norwegian text symbols distribution
Probabilities
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
r n t s l d k g m v f p b h j c w z x q

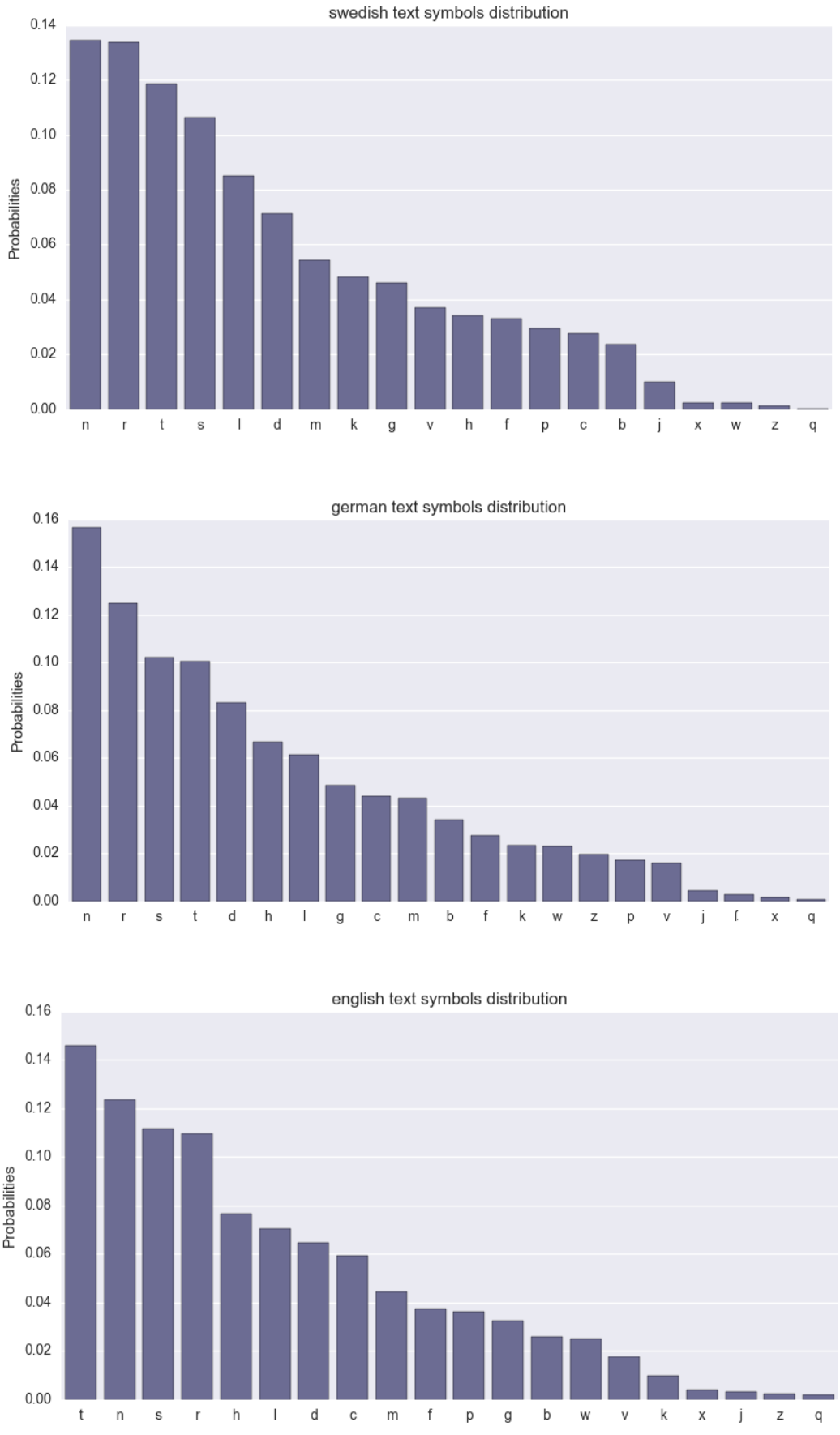


Рис. 1-15. Упорядоченное распределение символов в европейских языках

2. Распределение показателя Херста

Ранее в [1] было выяснено, что ряд расстояний между одинаковыми буквами в тексте (т.е. количество других символов, заключенных между выбранными последовательными двумя) стационарен и близок к ряду, образованному хаотической динамической системой. Распределение показателя Херста такого ряда унимодально и показывает антиперсистентность расположения символов в тексте.

Показатель Херста вычисляется следующим образом. Для данного временного ряда $b(t)$ строится ряд $x(t) = b(t+1) - b(t)$ его первых разностей и вводится скользящее среднее приростов по выборке длины k:

$$\bar{x}(t,k) = \frac{1}{k} \sum_{i=t-k+1}^{t} x(i).$$

Затем вычисляется накопленное отклонение от среднего (размах):

$$R(t,k) = \max_{j \le t} \left(\sum_{i=t-k+1}^{j} \left(x(i) - \bar{x}(t,k) \right) \right) - \min_{j \le t} \left(\sum_{i=t-k+1}^{j} \left(x(i) - \bar{x}(t,k) \right) \right).$$

Вычисляются также скользящая дисперсия рассматриваемого временного ряда по выборке длины k

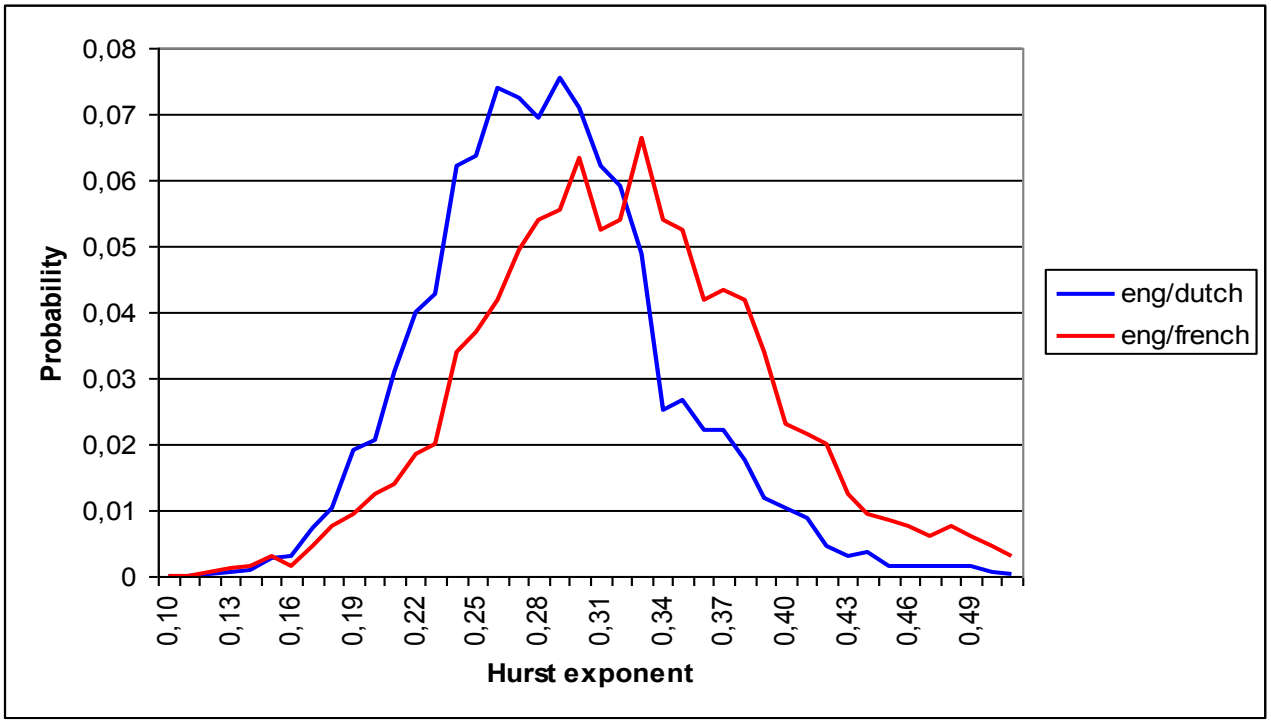
$$\sigma_x^2(t,k) = \frac{1}{k} \sum_{i=t-k+1}^{t} \left(x(i) - \bar{x}(t,k) \right)^2,$$

логарифм отношения размаха к шуму и его выборочное среднее:

$$\xi(t,k) = \ln\left(\frac{R(t,k)}{\sigma_x(t,k)} \right).$$

Показатель Херста $H_N(t)$ по выборке длины N на шаге t определяется как коэффициент регрессии величины $\xi(t,k)$ на логарифм $\ln N$ длины выборки. Это довольно затратная в вычислительном отношении процедура позволяет тем не менее с достаточной уверенностью различать тексты, написанные на языках из смешанных языковых групп, если длина окна перемешивания достаточна для построения асимптотики накопленного логарифмического размаха. Так, распределение показателя Херста для языков одной группы не меняется при их смешивании, а для смеси двух языков в отношении приблизительно 50/50 из разных групп наблюдается сдвиг распределения вправо, снижение моды и расширение носителя.

На рис. 16-17 приведены примеры смешивания языков из германской и романской групп. Важно, что сама по себе мода распределения не является надежным индикатором языка, так что гипотеза, выдвинутая в [2] относительно индикативных свойств этого статистического показателя для текста на одном языке, не подтвердилась на большом корпусе текстов. Тем не менее, этот индикатор можно использовать для анализа текста на однородность языка в случае использования одинакового алфавита.



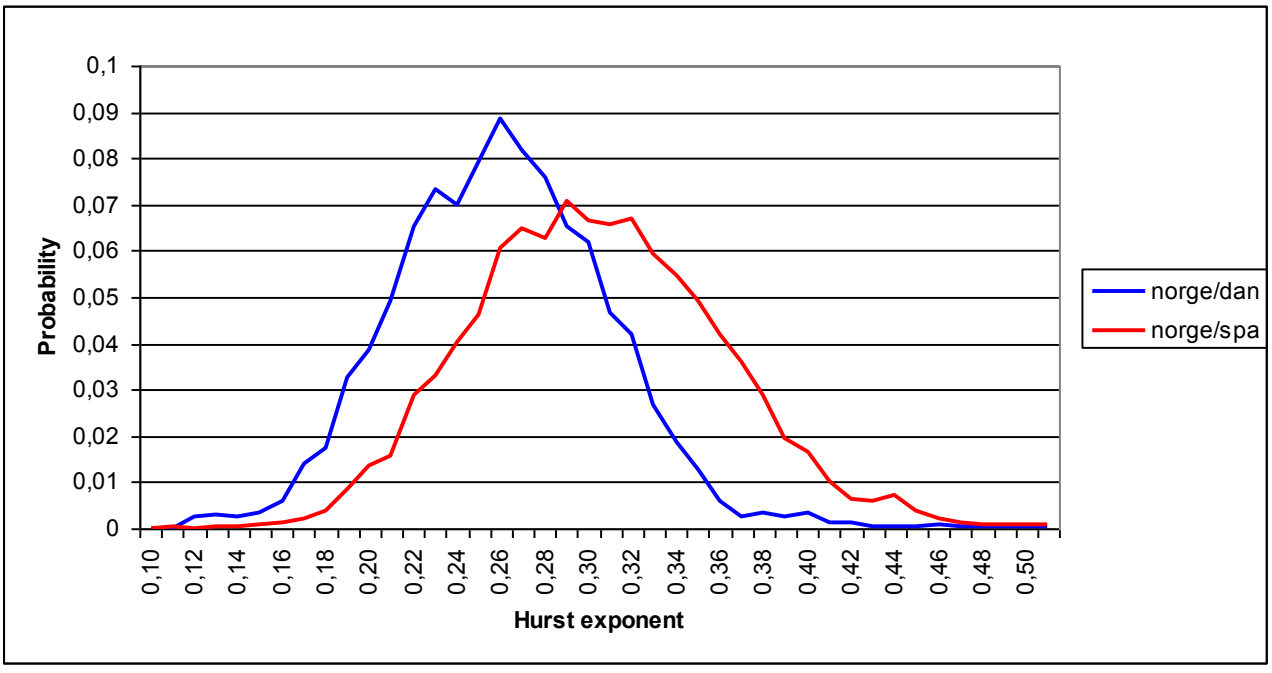


Рис. 16-17. Распределение показателя Херста ряда расстояний между буквой t при смешивании языков из одной и разных групп

С увеличением числа смешиваемых языков из разных групп распределение все более расширяется и его унимодальность теряется.

3. Наиболее вероятные трехбуквенные сочетания

Исследуем теперь вопрос о том, существуют ли индикативные буквосочетания, характерные для определенных языков. Тексты, напомним, исследуются без учета гласных букв.

Рассмотрим сначала те алфавиты, на которых изначально написаны тексты. Это нужно для того, чтобы проверить гипотезу о хорошем качестве классификации на очевидно различающихся текстах.

Берется смешанная выборка из 75 тысяч строк, каждая строка длиной 100 символов, по 5000 строк на каждый язык. Для классификации разобьем выборку на 3 части – на $2/3$ объема будем обучать наш алгоритм и на $1/3$ будем предсказывать класс (язык). Каждая строка перекодируется в матрицу векторов 3-х и 4-х грамм, в соответствующем столбце (например, для 3-граммы “mnl”) будет стоять единица у тех объектов (строк), которые эту 3-грамму содержат, в противном случае стоит 0. Всего различных n-грамм для этих языков получается достаточно много. Чтобы не усложнять модель, оставим только 200000 наиболее информативных. Для удобства хранения данных в памяти, воспользуемся разреженными матрицами Compressed Sparse Row (CSR) формата. Точность предсказания в таком случае составляет 98%.

Зависимость точности предсказания от длины самого фрагмента показана на рис. 18 – 19 и в табл. 1.

Табл. 1. Зависимость точности распознавания языка текста от длины фрагмента (число строк), «родной» алфавит

len	score
5	0.4788
10	0.668577777778
15	0.767111111111
20	0.820311111111
25	0.857422222222
35	0.9064
50	0.943511111111
75	0.972711111111
100	0.982133333333
200	0.994755555556
500	0.997955555556
1000	0.998844444444

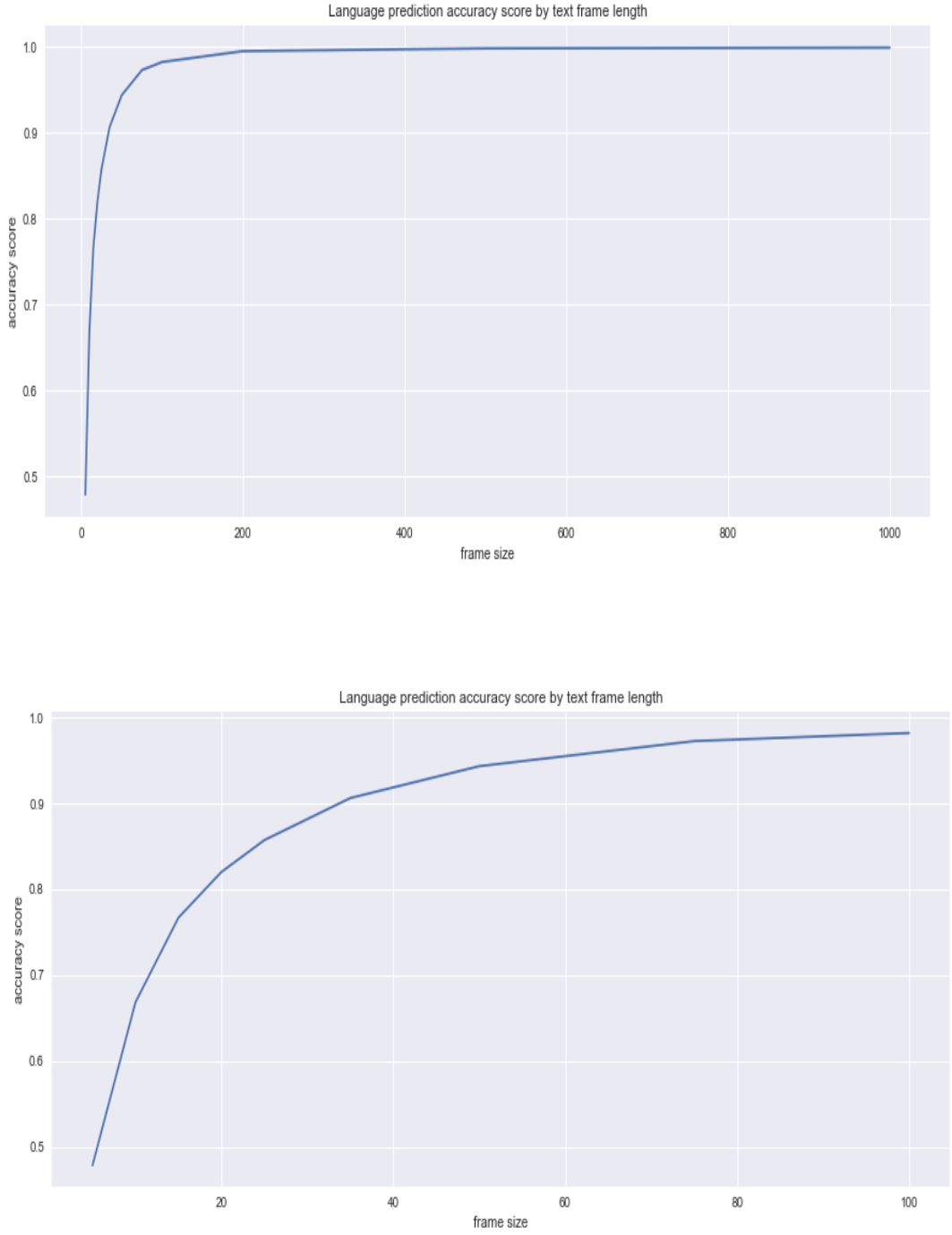


Рис. 18-19. Зависимость точности распознавания языка текста от длины фрагмента («родной» алфавит)

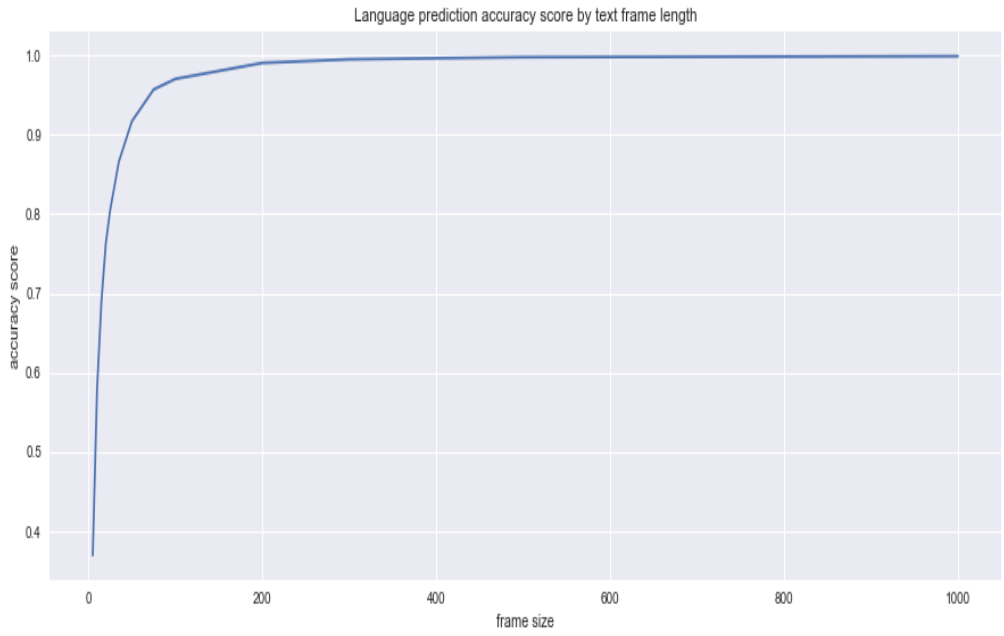
Таким образом, точность идентификация языка текста по относительно небольшим фрагментам весьма высока.

Теперь можно перейти к основной задаче – различения языка текста, написанного на одном алфавите (латиница). Построим аналогичные графики для этого случая (табл. 2, рис. 20-21).

Табл. 2. Зависимость точности распознавания языка текста
от длины фрагмента (число строк), один алфавит (латиница)

len	score
5	0.369955555556
10	0.580311111111
15	0.690133333333
20	0.762711111111
25	0.804711111111
35	0.865955555556
50	0.916844444444
75	0.956933333333
100	0.970266666667
200	0.9904
300	0.994977777778
500	0.997688888889
750	0.998311111111
1000	0.998933333333

Из сравнения табл. 1 и 2 следует, что на совсем малых выборках идентификация языка в латинском алфавите реализуется точнее, чем языка в «родном» алфавите, но с увеличением длины выборки ситуация меняется. Тексты, написанные на своем алфавите, определяются лучше, чем после транслитерации в латиницу, начиная с 500 строк.



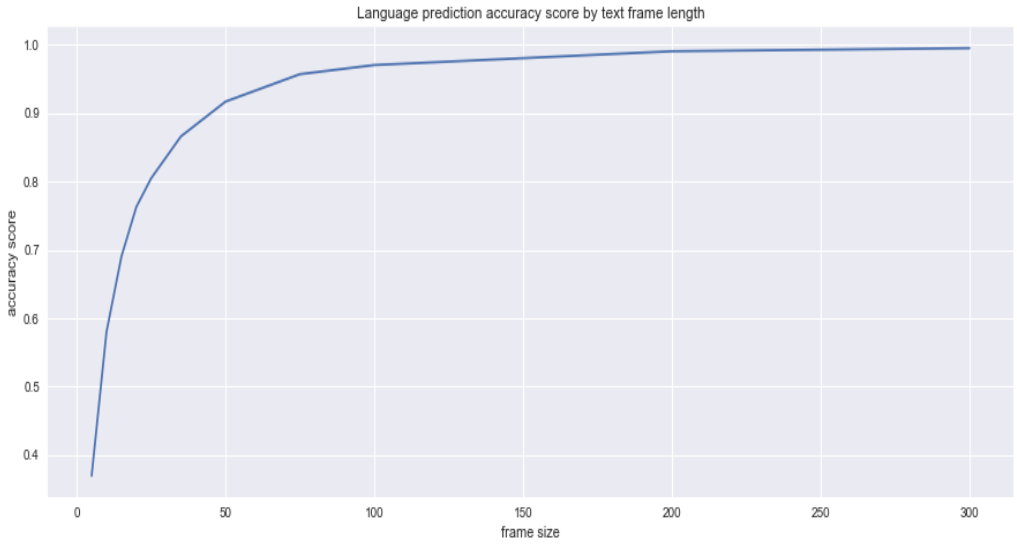


Рис. 20-21. Зависимость точности распознавания языка текста от длины фрагмента (один алфавит)

Количество n-грамм, на которых обучался классификатор для текстов на латинице, составило ~125000. Это означает, что таких признаков слишком много, чтобы можно было оценить их сравнительную эффективность визуально, поэтому интересно проверить, какие именно базовые n-граммы содержат в себе максимум информации о том или ином языке. Для этого обучим модель на разных значениях n максимального количества признаков и проверим зависимость качества определения языка от длины индикатора. Признаки при этом выбираются не случайным образом, а именно – K лучших, упорядоченных по частоте встречаемости в корпусе документов (объектов, строк). Далее будем рассматривать фреймы фиксированной длины 100 строк, на выборке с текстами в исходном алфавите качество определения языка для такой длины текстов составляет 0.982, для текстов в едином алфавите (в латинице) - чуть хуже, 0.971 (табл. 3, рис. 22).

Табл. 3. Зависимость точности распознавания от числа n-грамм

k-best	score
10	0.258577777778
25	0.380888888889
50	0.438577777778
100	0.48
250	0.711911111111
500	0.815066666667
1000	0.884977777778
2500	0.933911111111
5000	0.949866666667
10000	0.962711111111
25000	0.969022222222
50000	0.970577777778
100000	0.971111111111
125000	0.9712

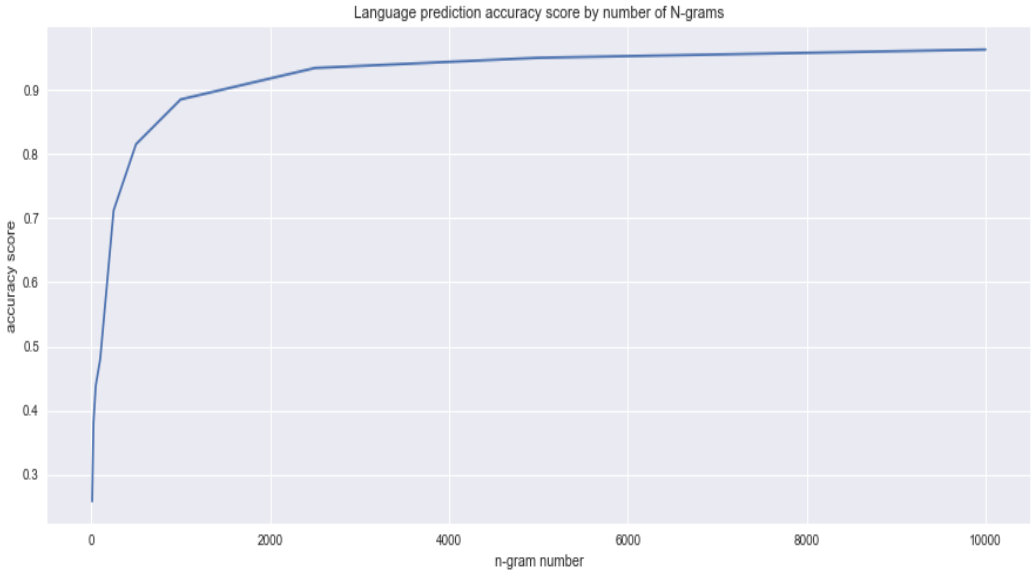


Рис. 22. Зависимость точности распознавания языка от числа *n*-грамм

Выделим теперь первые 20 *n*-грамм, которые оказались лучшими, и укажем, какие именно языки эти *n*-граммы описывают лучше всего в порядке убывания их эффективности.

- sch немецкий
- lctr румынский
- dll итальянский
- ctr румынский
- tncm румынский
- lct румынский
- wrd хорватский
- czn итальянский
- tnc румынский
- ths английский
- thr английский
- tht английский
- blgr румынский
- wdz датский
- chn немецкий
- scw голландский
- prz польский
- fth английский
- pntr румынский
- rlr румынский

Таким образом, на высоком уровне доверия с помощью небольшого количества *n*-грамм классифицируется примерно половина из рассмотренных

языков. Исследуем теперь вопрос о классификации языков на языковые группы, которых в нашем тесте пять:

1. славянская группа, западная подгруппа: польский, чешский;
2. славянская группа, южная подгруппа: сербский, хорватский, болгарский;
3. германская группа, северная подгруппа: датский, шведский, норвежский;
4. германская группа, западная подгруппа: немецкий, английский, голландский;
5. романская группа: итальянский, испанский, французский, румынский.

Такой подход предположительно позволит классифицировать тексты с большей точностью, а также уменьшить количество решающих n-грамм, максимально описывающих ту или иную группу (подгруппу). По аналогии с предыдущими пунктами и на тех же данных построим зависимость качества определения языковой группы фрагмента текста в зависимости от длины этого фрагмента и от количества решающих n-грамм (табл. 4, рис. 23).

Табл. 4. Зависимость точности идентификации группы от длины текста

length	score
5	0.5708
10	0.740666666667
15	0.825511111111
20	0.8696
25	0.904266666667
35	0.939066666667
50	0.9664
75	0.985422222222
90	0.9888
100	0.990311111111

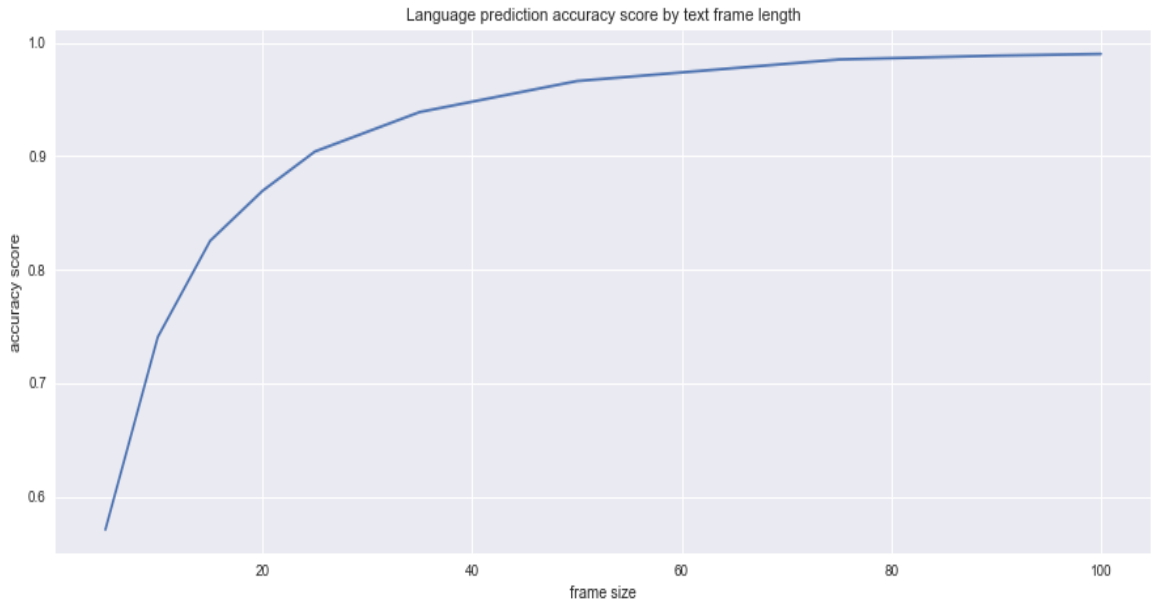


Рис. 23. Зависимость точности распознавания языковой группы от длины фрагмента текста (число строк)

Выяснилось, что языковая группа на тех же данных предсказывается гораздо точнее, чем сам язык. Зависимость точности идентификации от количества n-грамм показана в табл. 5 и на рис. 24.

Табл. 5. Зависимость точности идентификации группы от числа n-грамм

k-best	score
10	0.566577777778
25	0.684177777778
50	0.7524
100	0.843644444444
250	0.919244444444
500	0.943911111111
1000	0.964355555556
2500	0.980088888889
5000	0.985555555556
10000	0.9892
25000	0.989911111111
50000	0.990444444444
100000	0.9904
125000	0.990355555556

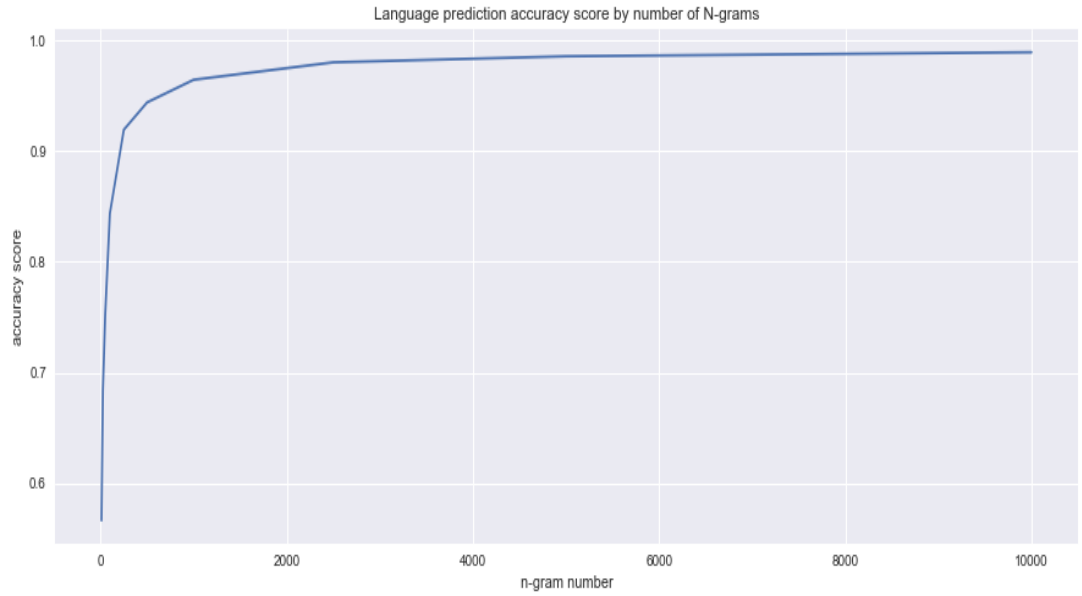


Рис. 24. Зависимость точности идентификации группы от числа n-грамм

Становится ясно, что для предсказания языковой группы требуется гораздо меньшее количество ключевых n-грамм, более того, начиная с какого-то момента (примерно 40-50 тыс. n-грамм) увеличение количества n-грамм только ухудшает качество предсказания. Это несколько неожиданный результат: оказывается, определенный квантиль распределения частот буквосочетаний работает лучше, чем распределение в целом. Более того, если построить полное распределение 3-грамм для большой совокупности текстов на определенном языке латинского алфавита с учетом гласных, то точность идентификации языка по близости распределения фрагмента текста к

соответствующему эталону окажется значительно хуже, чем для того же текста без гласных. Точность распознавания языка фрагмента по близости к полному распределению 3-грамм представлена в табл. 6.

Табл. 6. Точность распознавания языка фрагмента от его длины по близости к полному эталонному распределению 3-грамм

Длина фрагмента (число знаков)	**точность**
10000	0.101
25000	0.197
50000	0.360
100000	0.603
250000	0.773
500000	0.821
750000	0.829
1000000	0.872
1500000	0.835
2000000	0.845

Сравнивая результаты табл. 6 с табл. 2, видим, что распознавание языка по методу полного эталона на длине текста 2 млн знаков имеет ту же точность, что и метод наличия характерных буквосочетаний на длине всего лишь 35-40 строк (примерно 2 тыс. знаков).

Ниже приведены 20 лучших n-грамм для задачи идентификации языковой группы:

- sch германская группа, западная подгруппа
- wrd славянская группа, южная подгруппа
- stv славянская группа, южная подгруппа
- czn славянская группа, западная подгруппа
- prz славянская группа, западная подгруппа
- cht германская группа, западная подгруппа
- tht германская группа, западная подгруппа
- mnt романская группа
- wdz германская группа, западная подгруппа
- dll романская группа
- nsk германская группа, северная подгруппа
- scw славянская группа, южная подгруппа
- ctr романская группа
- cns романская группа
- ndr германская группа, западная подгруппа
- cws славянская группа, южная подгруппа
- chn германская группа, западная подгруппа
- lct романская группа
- lctr романская группа
- wsk славянская группа, южная подгруппа

В отличие от классификации отдельных языков, языковые группы полностью охвачены первыми двадцатью n-граммами.

Рассмотрим теперь бинарную классификацию текстов: насколько точно можно различить два отдельно взятых языка из разных языковых групп, при условии, что альтернатива известна точно. Например, рассмотрим разделение испанского и английского языков (табл. 7).

Табл. 7. Точность бинарной классификации в зависимости от длины текста (число строк)

length	score
5	0.784333333333
10	0.875266666667
15	0.9286
20	0.9492
25	0.965866666667
35	0.981133333333
50	0.9894
75	0.993866666667
90	0.9962
100	0.9974

Для двух языков суммарное количество n-грамм выходит чуть меньшим, чем для языковых групп – 100000 против 125000. По таблице качества классификации видно, что для распознавания одного языка из двух возможных с достоверностью более 95% , достаточны образцы текстов длиной всего 25 строк. Это в 4 раза меньше, чем для различения на том же уровне всех 15 языков. Рассмотрим качество классификации при различном количестве лучших n-грамм (табл. 8). Оказалось, что всего 10 ключевых n-грамм хватает, чтобы отличить английский язык от испанского с вероятностью ~93% и 500 n-грамм, чтобы точность различимости составляла ~99% (полная выборка - 99755 n-грамм).

Из 15-ти n-грамм 12 английских (первые 10 и последние 2) и 3 испанских.

Английские: thr, ths, nth, tht, sth, tth, fth, dth, thn, rth, wth, thc.

Испанские: dls, sdl, cnl.

Табл. 8. Точность бинарной классификации в зависимости от числа n-грамм

k-best	score
10	0.928533333333
25	0.956266666667
50	0.9714
100	0.979066666667
250	0.987266666667
500	0.991466666667
1000	0.993733333333
2500	0.996266666667
5000	0.996866666667
10000	0.997
25000	0.997333333333
50000	0.997333333333
90000	0.997266666667

Аналогично можно провести анализ точности бинарной классификации для других пар языков. Результаты оказываются близкими к представленным в табл. 8.

Заключение

Проведенное исследование показало, что для устойчивой идентификации языка текста с помощью распределения n-грамм сочетания согласных букв достаточно использовать относительно небольшое количество самых часто встречаемых буквосочетаний. Использование же полного распределения снижает точность идентификации. Добавление в полное распределение гласных еще более ухудшает точность.

При наличии дополнительных сведений, таких, как, например, выбор из двух возможностей, достаточное число n-грамм заметно снижается.

В дальнейшем предполагается распространить описанную методику на анализ других языковых групп.

Литература

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

2. 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.
URL: http://library.keldysh.ru/preprint.asp?id=2016-52