ЧЕБЫШЕВСКИЙ СБОРНИК
Том 23. Выпуск 2.

УДК 519.722 DOI 10.22405/2226-8383-2022-23-2-151-160

Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов

А. Г. Малашина, А. Б. Лось

Малашина Анастасия Геннадьевна — Национальный исследовательский университет «Высшая школа экономики» (г. Москва).
e-mail: amalashina@hse.ru
Лось Алексей Борисович — кандидат технических наук, доцент, Национальный исследовательский университет «Высшая школа экономики» (г. Москва).
e-mail: alos@hse.ru

Аннотация

При исследовании криптографических качеств алгоритмов защиты информации важным моментом является построение теоретических и экспериментальных моделей источников сообщений. В данной статье проводится статистический анализ свойств лексических и n-граммных моделей русского языка на основе новостного текстового корпуса. Создан специализированный корпус из новостных статей последних лет политической направленности, отражающий узкую область употребления языка. Составлены словари токенов и n-грамм, найдены величины покрытия этих словарей, а также значения энтропии. Проведена лемматизация исходного текстового корпуса и экстраполяция роста объёма словарей в зависимости от увеличения размера корпуса.

Ключевые слова: словари n-грамм, энтропия n-грамм, осмысленные тексты.

Библиография: 15 названий.

Для цитирования:
А. Г. Малашина, А. Б. Лось. Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов // Чебышевский сборник, 2022, т. 23, вып. 2, с. 151–160.

CHEBYSHEVSKII SBORNIK
Vol. 23. No. 2.

UDC 519.722 DOI 10.22405/2226-8383-2022-23-2-151-160

The construction and analysis of the Russian language models for a cryptographic algorithm research

A. G. Malashina, A. B. Los

Malashina Anastasia Gennad'evna — National Research University «Higher School of Economics» (Moscow).
e-mail: amalashina@hse.ru
Los Alexey Borisovich — candidate of technical sciences, assosiate professor, National Research University «Higher School of Economics» (Moscow).
e-mail: alos@hse.ru

Abstract

The article provides a statistical analysis of the properties of lexical and n-gram models of the Russian language based on the news text corpus. A specialized corpus of political news articles of recent years has been created, reflecting a narrow area of language use. The token and n-gram dictionaries are compiled, the coverage values are found, as well as the values of entropy. Lemmatization of the original text corpus and extrapolation of the dictionary volumes are performed.

Keywords: n-gram dictionaries, n-gram entropy, meaningful texts.

Bibliography: 15 titles.

For citation:
A. G. Malashina, A. B. Los, 2022, "The construction and analysis of the Russian language models for acryptographic algorithm research" , *Chebyshevskii sbornik*, vol. 23, no. 2, pp. 151–160.

1. Введение

Исследование вероятностного и статистического распределения слов и n-грамм естественного языка является предметом анализа во многих областях: лингвистике, теории игр, молекулярной биологии. Важную роль корпусный анализ языка играет в вопросах криптографической защиты информации, в том числе в вопросах эффективности ряда криптографических алгоритмов. При исследовании процедур восстановления отдельных участков сообщения по имеющейся информации о вариантах его знаков, основу анализа составляют создаваемые словари различных длин. В связи с этим, особое значение имеет изучение их статистических свойств, проверка полноты и адекватности используемого корпуса [6].

При исследовании языкового корпуса и составлении словарей, одним из основных вопросов является вопрос покрытия словарем всех возможных отрезков текста [15]. При этом проблема покрытия существенно усложняется для флективных языков, таких как французский и немецкий, и в особенности русский, по сравнению с аналитическими языками, такими как английский. Такие языки требуют большего объёма словаря для достижения необходимого покрытия [9].

В работе исследуются две языковые модели: лексическая и n-граммная. В лексической модели языка единицей анализа являются токены, то есть единицы текста, элементы раздельного написания. В n-граммной модели, являющейся частным случаем лексической, рассматриваются последовательности из n символов или слов.

2. Основной текст статьи

2.1. Языковой корпус

Корпус - собрание текстов в текстовой форме, используемое для исследования языка с использованием компьютерных технологий [4]. В данной работе был создан специализированный корпус русского языка, который отражает узкую область его употребления.

В качестве исходного материала для составления корпуса были использованы новостные статьи последних лет политической тематики. Эти тексты отражают срез состояния современного русского языка, включая разговорный, то есть составляемый корпус является синхроническим [4, 5].

После создания текстового корпуса осуществляется его нормализация, состоящая из следующих этапов: 1) удаление html-тегов и переформатирование в *.txt; 2) перекодировка; 3) удаление всех сокращений, кроме аббревиатур; 4) удаление имён нарицательных ; 5) фильтрация текста (удаление всех символов, кроме «а-я», «.», «,», « », приведение к нижнему

регистру); 6) удаление двойных пробелов, повторяющихся точек и запятых, пробелов перед точками и запятыми.

Метаданные в корпусе отсутствуют, так как их наличие не принципиально для дальнейших целей использования данного корпуса.

Созданный корпус должен удовлетворять двум основным критериям: полноте и репрезентативности. Полнота корпуса обуславливается покрытием этого корпуса. Оптимизация покрытия зависит от задач, для которых создаётся этот корпус и словари. Во-первых, покрытие зависит от объёма текстового корпуса, который используется для построения словарей, но с определенного момента эта зависимость становится гораздо менее выраженной, поэтому возможна экстраполяция логарифмическими функциями. Например, для английского языка рост объёма словаря существенно замедляется при размере корпуса от 30 до 50 млн. слов. Во-вторых, оптимальный размер корпуса зависит от источников и новизны данных [9]. В целом, корпус считают насыщенным, когда с увеличением объёма корпуса прекращается резкий рост новых слов [3].

Репрезентативность — это способность корпуса адекватно отражать специфику выбранной предметной области [4].

На основе собранного новостного текстового корпуса в соответствии с рассматриваемой языковой моделью создаются словари, которые впоследствии подвергаются статистическому исследованию.

2.2. Лексическая модель

Лексические модели языка представляют особый интерес для систем распознавания речи. Актуальность вопроса максимизации покрытия связана с тем, что каждое неизвестное слово (называемое также out-of-vocabulary или OOV) создаёт очередную ошибку в распознавании текущего слова. Более того, каждая такая ошибка способна породить ошибку распознавания следующего слова, создавая «волновой эффект» OOV-слов.

В рамках лексической модели языка генерируются словари, единицами которых являются токены, то есть единицы текста как элементы раздельного написания. В таблице ниже представлены значения размеров словарей, создаваемых на основе корпусов различного объёма.

Корпус, симв.	Объем словаря	Эмпирическое покрытие	Теоретическое покрытие
10^4	836	0.45	20.1
10^5	5449	2.74	28.34
10^6	31895	13.78	39.68
10^7 с нарицат.	163091	70.63	48.29
10^7 без нарицат.	125162	64.45	44.98
10^8	*180000*	-	*59.28*
10^9	*230000*	-	*68.84*

Поскольку скорость роста объёма словарей в зависимости от размера корпуса близка к скорости роста логарифмической функции, проводится следующая экстраполяция:

$$21910.5 \cdot \ln(3.434 \cdot 10^{-6} \cdot x). \quad (1)$$

Прогнозируемые значения объёма словарей обозначены в таблице курсивом. Графически данная зависимость отражена на рисунке 1.

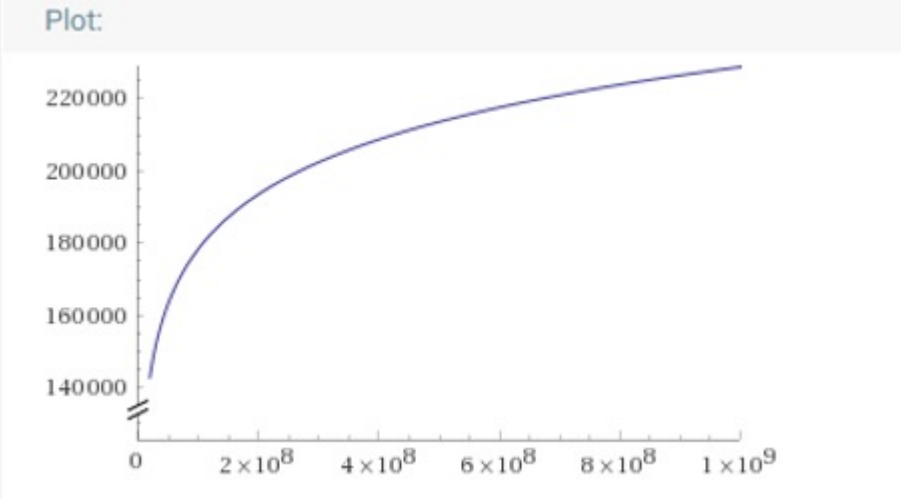

Рис. 1: Размер словаря в зависимости от объёма корпуса.

2.3. Закон Ципфа

Для проверки качества и естественности созданного русскоязычного корпуса (объёмом 10^7 символов) проводится проверка соответствия закону Ципфа. В соответствии с данным законом если все слова в корпусе упорядочить по убыванию частоты их встречаемости, то частота использования слов окажется обратно пропорциональной их порядковому рангу [12, 13]. Результаты представлены на рисунке 2.

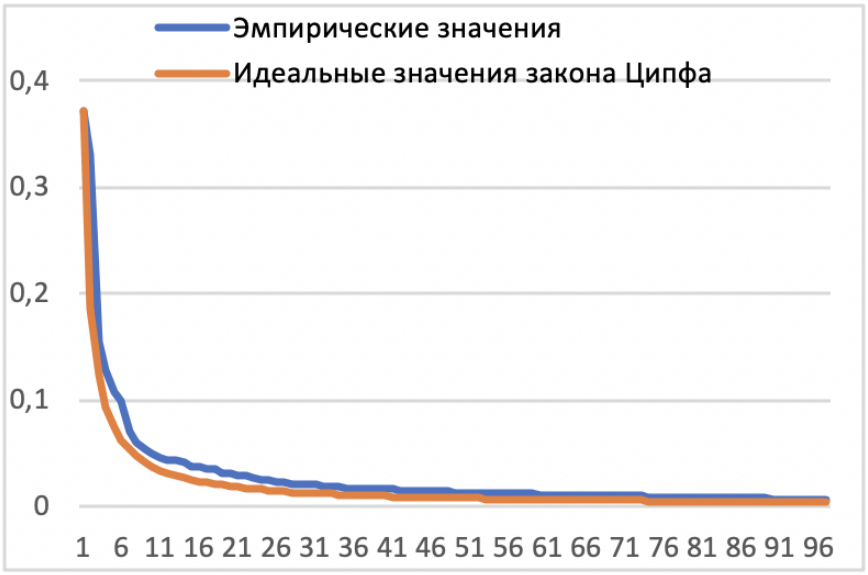

Рис. 2: Закон Ципфа.

Как наглядно демонстрирует график, эмпирически полученные значения лишь незначительно отклоняются от идеальных значений закона Ципфа. Поэтому можно заключить, что собранный корпус, в целом удовлетворяет данному закону и является достаточно естественным и полным в рамках лексической модели.

2.4. Покрытие

Для оценки полноты рассматриваемого корпуса требуется оценить величину покрытия его словаря. Существуют различные подходы к оценке покрытия. В данном исследовании были рассмотрены эмпирическая и теоретическая оценки.

Эмпирическая оценка проводится с помощью тестового корпуса, не являющегося частью проверяемого. Величина покрытия — это доля токенов тестового корпуса, покрытая словарём исходного корпуса.

Теоретическая оценка основывается на количестве уникальных токенов словаря, то есть единиц словаря, встречающихся только один раз:

$$\tilde{N} = \frac{N}{1 - \frac{n}{N}}, \tag{2}$$

где N – исходный объём словаря, n – число токенов, встречающихся один раз.

Предполагается, что полученная таким образом величина – это приблизительный размер полного словаря. На основании теоретически найденного объёма словаря оценивается величина покрытия как отношение практической величины объема к его теоретической величине. При этом величина покрытия выражается в процентах.

Значения эмпирического и теоретического покрытия указаны в таблице 1. С увеличением объёма корпуса величина покрытия возрастает. Для небольших корпусов теоретически найденное покрытие значительно превышает экспериментальное. Такое различие вскрывает недостатки подхода, основанного на количестве однократно встречаемых единиц словаря, особенно для корпусов небольшого объёма. Для больших корпусов эмпирическое покрытие выше теоретического. Данное обстоятельство может быть объяснено тем, что среди однократно встречающихся токенов в словаре могут присутствовать слова, написанные с опечатками, редкие имена и фамилии, специфические названия, узкоспециализированная лексика и редко употребляемые слова. В качестве примера, демонстрирующего подобный недостаток теоретической оценки покрытия, можно привести Брауновский корпус английского языка. Более половины слов в данном корпусе встречаются лишь однажды. Следовательно, относительно теоретической оценке его покрытие составляет менее 50% несмотря на то, что объём корпуса около 1 млн. слов [4].

2.5. Лемматизация

Лемматизация - процесс обработки языкового корпуса, в результате которого все входящие в него слова приводятся к словарной форме [4].

Для исследуемых корпусов проводится лемматизация, после чего оценивается величина покрытия. Сравнительные результаты приведены в таблице.

Корпус	Покрытие (оригинал)	Покрытие (лемматизация)
10^6	13.78	16.16
10^7 с нарицат.	70.63	76.45
10^7 без нарицат.	64.45	66.77

Результаты наглядно демонстрируют, что величины покрытия лемматизированных корпусов выше покрытия оригиналов. Это связано с тем, что лемматизация значительно снижает флективность языка, приближая его к аналитическому [9]. Например, в исследуемом корпусе присутствует слово «репрезентативные», а в тестовом корпусе только слово «репрезентативный». То есть одно и то же слово присутствует в корпусах в разной форме. Но с точки зрения автоматической обработки, точно сопоставляющей токены, это разные слова. Это снижает величину покрытия. Лемматизация устраняет данный недостаток и снижает объём словаря.

Процесс лемматизации может оказаться крайне полезен, например, для генерации словарей на основе корпусов. Но для некоторых других задач, таких как распознавание речи или восстановление открытого текста без определения ключа шифрования, словаря (в более общем смысле), содержащего только словарные формы, может оказаться недостаточно. В таком случае расширение словаря может производиться в два этапа: с помощью лемматизации, а затем процесса, обратного ей. Например, в корпусе встречается только слово «языковые». Леммой данного слова является слово «языковой». В словарь могут добавляться все орфографические формы данного слова: «языковая», «языковое», «языковых», «языковым» и т.д. Более того, словарь может быть расширен ещё больше, если добавлять в него также однокоренные токены других частей речи, например, «язык». Но такой подход требует наличия промежуточного автоматизированного блока, умеющего производить орфографические изменения слов.

2.6. Ключевые слова

На основе корпуса объёмом 1 млн. символов определяются ключевые слова, то есть слова, встречающиеся чаще всего. Результаты представлены в двух вариантах: для исходного и лемматизированного корпуса. Топ-10 слов приведен в таблице ниже:

№	Лемматизированный корпус	Количество	Оригинальный корпус	Количество
1	Россия	663	России	420
2	страна	76.45	Сирии	290
3	политический	66.77	США	270
4	Сирия	16.16	время	228
5	год	76.45	страны	208
6	человек	66.77	Россия	173
7	российский	16.16	политической	160
8	один	76.45	власти	146
9	много	66.77	против	142
10	время	66.77	РФ	133

Ключевые слова подчеркивают специфичность и специализированность корпуса, собранного на основе современных новостных статей на тему политики.

2.7. N-граммная модель

В n-граммной модели единицами корпуса (словаря) считаются последовательности из n символов. Лексические модели являются подмножеством n-граммных моделей языка.

Вопросы n-граммного покрытия представляют интерес в системах распознавания речи для максимизации производительности системы. Но, особое значение, n-граммные модели имеют для вопросов криптографии, так как в шифрованном тексте границы между словами неизвестны, и подбор в этом случае по лексическому словарю невозможен или крайне затруднителен [11]. Если какая-либо n-грамма отсутствует в словаре, то языковая модель может опираться на n-граммы более низкого порядка, но они могут оказаться неуместны для текущей задачи. Именно поэтому ошибки распознавания гораздо чаще происходят в рамках n - граммной модели языка [10].

Анализ покрытия словаря n-грамм усложняется из-за значительно меньшей частоты n-грамм по сравнению с наименее частыми словами в словаре. По оценкам сборника североамериканских новостных деловых статей (NAB), который на данный момент является самым тщательно исследованным языковым корпусом, чтобы оптимизировать охват биграмм (два слова), требуется корпус объёмом от 100 до 200 млн. слов. Собрать текстовой корпус такого

объема представляется довольно трудной задачей. А с увеличением n проблема оптимизации покрытия только ухудшается.

Ещё больше осложняет ситуацию существующее потенциальное взаимодействие между покрытием n-грамм и эволюцией языка. Накопление слов из соответствующего источника, очевидно, занимает время, в течение которого языковые шаблоны могут меняться, ухудшая адекватность более старых данных. Рассматривая язык как нестационарный стохастический источник, Розенфельд постулировал следующий принцип: никогда нельзя определить одновременно и степень, и временные рамки языкового явления. Как следствие, он пришел к выводу, что невозможно обнаруживать преходящие и редкие лингвистические события [15].

На основе собранных корпусов проводится оценка покрытия и подсчитывается энтропия n-грамм длиною i символов [1, 2, 7, 14]:

$$H_i = \frac{log_2 N_i}{i}. \tag{3}$$

где i - длина n-граммы, а N_i - количество n-грамм в слове длиною i символов. Рассматриваются n-граммы длиною 10, 15, 20 и 25 символов. Результаты исследования приведены в таблице и на рисунке 3.

Длина n-граммы	Объем словаря		Энтропия	
	10^6	10^7	10^6	10^7
10	795840	6217191	1.96	2.26
15	955193	9482897	1.32	1.55
20	983828	10372296	0.99	1.17
25	990430	10629589	0.80	0.93

Значения энтропии n-грамм близки к реальным значениям для русского языка [8].

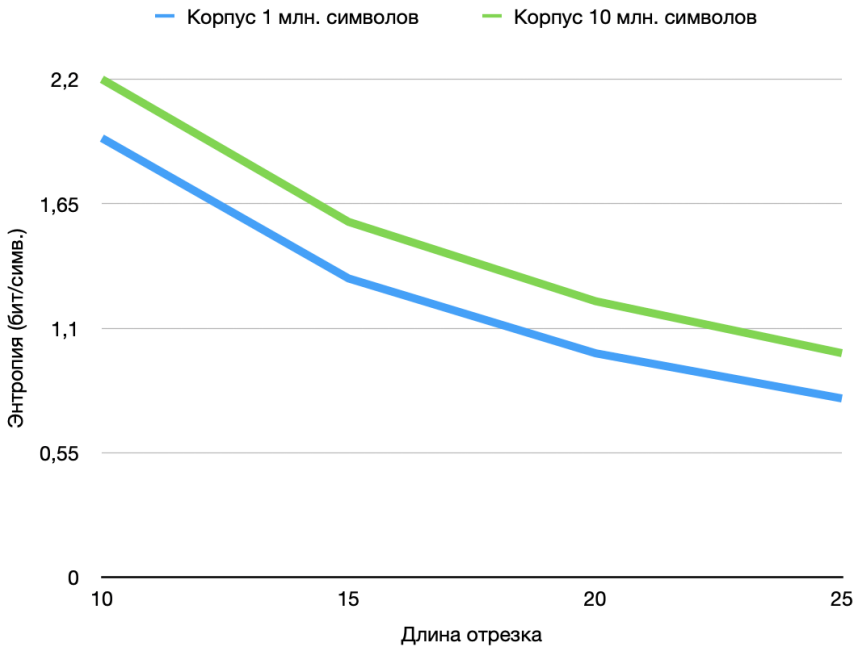

Рис. 3: Энтропия n-грамм.

Величины покрытия представлены в таблице ниже.

Длина n-граммы	Эмпирическое покрытие		Теоретическое покрытие	
	10^6	10^7	10^6	10^7
10	4.32	39.71	11.27	19.95
15	1.10	12.03	3.15	7.21
20	0.28	3.12	1.30	3.27
25	0.07	0.84	0.79	2.07

Как видно, покрытие словаря n-грамм значительно ниже лексического покрытия. Это подтверждает тот факт, что исследование n-граммной модели гораздо более сложное, а оптимизация покрытия словаря требует сверхбольшого языкового корпуса.

3. Заключение

В данной статье рассмотрены две основные модели естественного языка, которые были исследованы с помощью созданного текстового корпуса, основанного на новостных статьях последних лет. Рассмотрены различные подходы к определению величины покрытия текстов. Для лексической модели языка был проверен закон Ципфа, оценивающий естественность языкового корпуса, а также проведена лемматизация корпуса, подчеркивающая высокую флективность русского языка. Найдены значения информационной энтропии n-грамм ограниченной длины. Построена экстраполяционная модель дальнейшего изменения объема словарей.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

1. Алферов А. П., Зубов А. Ю., Кузьмин А. С., Черемушкин А. В., Основы криптографии: учебное пособие. 3-е изд., испр. и доп. // М.: Гелиос АРВ, 2005. – 408 с.

2. Бабаш А. В., Шанкин Г. П., Криптография, Москва: СОЛОН-ПРЕСС, 2007.

3. Викторов А. Б., Грамницкий С. Г., Гордеев С. С., Ескевич М. В. и Климина Е. М. Универсальная методика подготовки компонентов обучения систем распознавания речи // Речевые технологии, pp. 39-56, Февраль 2009.

4. Волосатова Т. М., Информатика и лингвистика: учеб. пособие, Волосатова Т. М. и Чичварин Н. В. // ИНФРА-М, 2018, 196 с.

5. Кипяткова И. С. Исследование статистических n-граммных моделей языка для распознавания слитной русской речи со сверхбольшим словарем // Анализ разговорной русской речи , Санкт-Петербург, 2010.

6. Малашина А. Г. Алгоритм восстановления отдельных частей сообщения по информации о возможных значениях его знаков, Материалы конференции // Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского. – Москва, 2019. С. 215-217.

7. Шеннон К., Работы по теории информации и кибернетике // М: Издательство иностранной литературы, 1963.

8. Яглом А. М., Яглом И. М., Вероятность и информация: 3-е изд., испр. и доп. // М: издательство «Наука», 1973, 236-290 с.

9. Bellegarda J. R. Robustness in Statistical Language Modeling // Robustness in Language and Speech Technology , Springer Science+Business Media Dordrecht, 2001, pp. 104-106.

10. Chase L., Rosenfeld R., Ward W. Error-responsive modifications to speech recognizers: negative n-grams // Third International Conference on Spoken Language Processing, Yokohama, 1994.

11. Florencio D., Herley C. A Large-Scale Study of Web Password Habits // Proceeds of the International World Wide Web Conference Committee, 2015.

12. Gelbukh A., Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language // Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2001.

13. Kechedzhy K. E. Rank distributions of words in additive many-step Markov chaons and the Zipf law // Phys. Rev. E. – 2005 – Vol. 72.

14. Massey J. Guessing and entropy // Proceedings of 1994 IEEE International Symposium on Information Theory. IEEE. p. 204.

15. Rosenfeld R. Optimizing lexical and n-gram coverage via judicious use of linguistic data // Proceedings of the Fourth European Conference on Speech Communication and Technology – Madrid, 1995.

REFERENCES

1. Alferov A. P., Zubov A. Ju., Kuz'min A. S. & Cheremushkin A. V. 2005, "Fundamentals of cryptography: textbook: 3rd ed., ISPR. and add."["Osnovy kriptografii: uchebnoe posobie: 3-e izd., ispr. i dop."], *Gelios ARV*, p. 408.

2. Babash A. V. & Shankin G. P. 2007, "Cryptography" ["Kriptografija"], *SOLON-PRESS*.

3. Viktorov A. B., Gramnickij S. G., Gordeev S. S., Eskevich M. V. & Klimina E. M. 2009, "The universal method for preparing speech recognition system training components" ["Universal'naja metodika podgotovki komponentov obuchenija sistem raspoznavanija rechi"], *Rechevye tehnologii*, pp. 39-56.

4. Volosatova T. M. & Chichvarin N. V. 2018, "Computer science and linguistics: textbook" ["Informatika i lingvistika: ucheb. posobie"], *INFRA-M*, p.196.

5. Kipjatkova I. S. 2010, "Research of statistical N-gram language models for recognition of merged Russian speech with a super-large dictionary" ["Issledovanie statisticheskih n-grammnyh modelej jazyka dlja raspoznavanija slitnoj russkoj rechi so sverhbol'shim slovarem"], *Analiz razgovornoj russkoj rechi*, Sankt-Peterburg.

6. Malashina A. G. 2019, "The algorithm for recovering discrete message parts based on information about possible values of its characters. Proc." ["Algoritm vosstanovlenija otdel'nyh chastej soobshhenija po informacii o vozmozhnyh znachenijah ego znakov. Materialy konferencii"], *Mezhvuzovskaja nauchno-tehnicheskaja konferencija studentov, aspirantov i molodyh specialistov imeni E. V. Armenskogo*, Moscow, pp. 215-217.

7. Shennon C. E. 1963, "Works on information theory and Cybernetics" ["Raboty po teorii informacii i kibernetike"], *Izdatel'stvo inostrannoj literatury*.

8. Jaglom A. M. & Jaglom I. M. 1973, "Probability and information: 3rd ed., cor. and exp." ["Verojatnost' i informacija: 3-e izd., ispr. i dop."], **Nauka**, pp. 236-290.

9. Bellegarda J. R. 2001, "Robustness in Statistical Language Modeling", *Robustness in Language and Speech Technology , Springer Science+Business Media Dordrecht*, pp. 104-106.

10. Chase L., Rosenfeld R. & Ward W. 1994, "Error-responsive modifications to speech recognizers: negative n-grams", *Third International Conference on Spoken Language Processing.*

11. Florencio, D. & Herley, C. 2007, "A Large-Scale Study of Web Password Habits", *Proceeds of the International World Wide Web Conference Committee.*

12. Gelbukh A. & Sidorov G. 2001, "Zipf and Heaps Laws' Coefficients Depend on Language", *Conference on Intelligent Text Processing and Computational Linguistics.*

13. Kechedzhy K. E., Usatenko K. E. & V. A. Yampol'skii 2005, "Rank distributions of words in additive many-step Markov chaons and the Zipf law", *Phys. Rev. E.*, vol. 72.

14. Massey J. 1994, "Guessing and entropy", *Proceedings of 1994 IEEE International Symposium on Information Theory*, p. 204.

15. Rosenfeld R. 1995, "Optimizing lexical and n-gram coverage via judicious use of linguistic data", *Proceedings of the Fourth European Conference on Speech Communication and Technology.*

Получено 30.09.2020
Принято в печать 22.06.2022