ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
МАТЕМАТИКИ
Том 23 Выпуск 1
2016

Секция **«Прикладная вероятность и статистика»**

ВИЛЬБОА Н. В.*, ***ЛОСЬ А. Б.****, ***МИРОНКИН В. О****

ОБ ИССЛЕДОВАНИИ ИНФОРМАЦИОННЫХ ХАРАКТЕРИСТИК ЕСТЕСТВЕННЫХ ЯЗЫКОВ

§ 1. Введение

Вероятностные и статистические характеристики естественных языков находят важное применение в различных областях знаний и приложениях, о чем свидетельствует большое число публикаций по этой тематике [5, 9–15], в том числе за последние годы.

В статье рассматриваются задачи, связанные с нахождением оценок ряда информационных характеристик (энтропия, распределение m-грамм, слогов и словосочетаний), использование которых позволяет строить прогнозные модели развития естественных языков. Проведены экспериментальные исследования текстов на русском, английском, немецком, французском и грузинском языках, классифицированных как по временным интервалам XVIII, XIX, XX и XXI веков, так и по стилям: художественный, научный, официально-деловой (политический). Каждая анализируемая группа данных содержала текстовый материал объемом 10^7 символов. Тексты выбирались из различных источников во избежание искажений, связанных с авторскими особенностями. Данные по XVIII, XIX векам на русском языке были дополнены текстами, написанными в дореволюционной орфографии, а современные материалы — текстами, взятыми из сети «Internet». Для каждой из описанных групп текстов были вычислены значения оценок энтропии и избыточности, а также приведены сводные таблицы и построены графики, иллюстрирующие поведение шаговой энтропии при увеличении глубины зависимости букв в тексте.

Приведем ряд используемых понятий и определений.

§ 2. Основные понятия и определения

Учитывая, что в лексике число элементарных единиц практически не ограничено, а в области фонетики в принципе трудно выделить дискретные единицы, исследование было ограничено изучением текстовых материалов.

*Москва, Национальный исследовательский университет «Высшая школа экономики»

Пусть задан произвольный конечный алфавит, на котором определена дискретная вероятностная схема:

$$A = \begin{pmatrix} a_1 & \dots & a_k \\ p(a) & \dots & p(a_k), \end{pmatrix}$$

где a_i — i-й исход вероятностной схемы A с вероятностью $p(a_i)$: $p(a_1) + p(a_2) + \dots + p(a_k) = 1,\ 0 < p(a_i) < 1$.

Пусть далее на основе вероятностной схемы определены объединенные вероятностные схемы $A^m,\ m \in \mathbf{N}$.

Замечание. Наиболее адекватной моделью для описания свойств естественных языков является модель марковского источника переменного порядка [9, 10].

В рамках проводимых исследований основной целью было описание характера изменения энтропии в течение анализируемых временных периодов, поэтому была использована более простая модель марковского источника сообщений фиксированного порядка [1–4, 6–8, 16]. В каждом из проведенных экспериментов значение порядка m, менялось от 1 до 10.

Определение. Дискретный стационарный источник $(A, \vec{p})$ называется марковским источником порядка m, если для любого $l \geqslant m$ и любой последовательности $(a_1, \dots, a_l)$ букв алфавита A справедливо равенство $p(a_l/a_1, \dots, a_{l-1}) = p(a_l/a_{l-m+1}, \dots, a_{l-1})$.

Определение. Шаговой энтропией марковского источника порядка называется величина:

$$H^{(m)} = -\sum_{(a_{i_1},\dots,a_{i_m}) \in A^m} p(a_{i_1}, \dots, a_{i_m}) \log_a p(a_{i_m}/a_{i_1}, \dots, a_{i_{m-1}}). \qquad (1)$$

Определение. Энтропией марковского источника на один знак называется величина

$$H_m = -\frac{1}{m} \sum_{(a_1,\dots,a_m) \in A^m} p(a_1, \dots, a_m) \log_a p(a_1, \dots, a_m).$$

Определение. Величина $\lim_{m\to\infty} H_m = \lim_{m\to\infty} H^{(m)} = H \geqslant 0$ называется энтропией марковского источника.

Определение. Избыточностью источника сообщений называется величина

$$R_m = 1 - \frac{H_m}{H_{\max}}, \qquad (2)$$

где $H_{\max} = \log_a |A|$.

Замечание. Для оценивания энтропии использовалось соотношение (1), в котором $a = 2$ (энтропия вычислялась в битах).

§ 3. Описание алгоритма

Представим шаговую энтропию марковского источника (1) в следующем виде:

$$H^{(m)} = -\sum_{(a_{i_1},\ldots,a_{i_m})\in A^m} p(a_{i_1},\ldots,a_{i_m}) \log_2 \frac{p(a_{i_1},\ldots,a_{i_m})}{p(a_{i_1},\ldots,a_{i_{m-1}})}. \tag{3}$$

В качестве оценки вероятности k-грамм $p(a_{i_1},\ldots,a_{i_k})$, $k \in \mathbf{N}$, будем использовать частоту встречаемости соответствующих k-грамм.

Через $N(a_{i_1},\ldots,a_{i_k})$ обозначим количество появлений k-граммы $(a_{i_1},\ldots,a_{i_k})$ в анализируемом объеме текстов, а через $N_k(l) = l-(k-1)$ — общее количество k-грамм заданного алфавита в анализируемом объеме текстов длины l.

С учетом (3) в качестве оценки энтропии будем использовать величину

$$\begin{aligned}
\widetilde{H}^{(m)} &= \frac{1}{N_m(l)} \sum_{(a_{i_1},\ldots,a_{i_m})\in A^m} N(a_{i_1},\ldots,a_{i_m}) \\
&\qquad \times \log_2 \left(\frac{N(a_{i_1},\ldots,a_{i_m})}{N_m(l)} \frac{N_{m-1}(l)}{N(a_{i_1},\ldots,a_{i_{m-1}})} \right) \\
&= -\frac{1}{N_m(l)} \sum_{(a_{i_1},\ldots,a_{i_m})\in A^m} N(a_{i_1},\ldots,a_{i_m}) \\
&\qquad \times \log_2 \left(\frac{N(a_{i_1},\ldots,a_{i_m})}{N(a_{i_1},\ldots,a_{i_{m-1}})} \frac{N_{m-1}(l)}{N_m(l)} \right) \\
&= -\frac{1}{N_m(l)} - \frac{1}{N_m(l)} \sum_{(a_{i_1},\ldots,a_{i_m})\in A^m} N(a_{i_1},\ldots,a_{i_m}) \\
&\qquad \times \left(\log_2 \frac{N(a_{i_1},\ldots,a_{i_m})}{N(a_{i_1},\ldots,a_{i_{m-1}})} + \log_2 \frac{N_{m-1}(l)}{N_m(l)} \right)
\end{aligned} \tag{4}$$

Из соотношения (2) получим оценку избыточности:

$$\begin{aligned}
\widetilde{R}_m &= 1 + \frac{\log_{|A|} 2}{N_m(l)} \sum_{(a_{i_1},\ldots,a_{i_m}\in A^m} N(a_{i_1},\ldots,a_{i_m}) \\
&\qquad \times \left(\log_2 \frac{N(a_{i_1},\ldots,a_{i_m})}{N(a_{i_1},\ldots,a_{i_{m-1}})} + \log_2 \frac{N_{m-1}(l)}{N_m(l)} \right)
\end{aligned} \tag{5}$$

Для вычисления значений $\widetilde{H}^{(m)}$, $\widetilde{R}_m$ согласно выражениям (4), (5) требуется найти все значения $N(a_{i_1}\ldots,a_{i_k})$, где $(a_{i_1}\ldots,a_{i_k}) \in A^m$, $m \in \overline{1,10}$.

Замечание. $N(a_{i_1} \ldots, a_{i_k})$, может быть оптимизировано за счет использования эффективных алгоритмов, например, [13, 15].

Для обработки анализируемых текстов был проведен ряд предварительных преобразований:

1. удаление данных о названии, авторе, самой публикации, издателе и т. п.;

2. удаление заголовков (частей, глав, разделов, подразделов и т. д.) и нечитаемых последовательностей символов;

3. удаление графической информации.

Для каждого языка был задан набор возможных символов, включающий алфавит данного языка, пробел, знаки пунктуации. Символы, не вошедшие в указанный набор, были исключены.

§ 4. Основные результаты

В параграфе представлены таблицы, содержащие полученные экспериментальные значения избыточности $\widetilde{R}_{10}$ и шаговой энтропии $\widetilde{H}^{(m)}$ в зависимости от величины параметра m для каждой группы текстов исследуемых языков. Точность вычислений составляла 10^{-3}.

Приведем сначала результаты экспериментов для русского языка.

Таблица 1. Русский язык. XVIII век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
Современная орфография										
худ.	4,111	3,508	2,830	2,177	1,525	1,175	0,892	0,653	0,440	0,893
офиц.-дел.	4,043	3,450	2,795	2,252	1,756	1,284	0,902	0,626	0,432	0,891
научн.	4,058	3,401	2,659	2,075	1,525	1,059	0,725	0,494	0,334	0,918
Дореволюционная орфография										
научн.	4,002	3,258	2,465	1,811	1,248	0,823	0,556	0,392	0,288	0,928

Таблица 2. Русский язык. XIX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
Современная орфография										
худ.	4,308	3,552	2,749	2,235	1,820	1,435	1,239	0,939	0,683	0,841
офиц.-дел.	4,004	3,146	2,422	1,966	1,587	1,269	0,914	0,713	0,455	0,836
научн.	4,196	3,45	2,693	2,125	1,644	1,335	1,017	0,885	0,517	0,828
Дореволюционная орфография										
худ.	3,997	3,403	2,644	2,063	1,491	0,995	0,635	0,403	0,255	0,936
офиц.-дел.	4,070	3,305	2,549	1,754	1,124	0,675	0,406	0,252	0,161	0,960
научн.	3,995	3,219	2,377	1,668	1,099	0,698	0,449	0,295	0,196	0,951

Таблица 3. Русский язык. XX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
Современная орфография										
худ.	4,308	3,652	2,749	2,335	1,620	1,385	1,179	0,839	0,583	0,865
офиц.-дел.	4,122	3,334	2,249	1,840	1,136	1,077	0,862	0,490	0,316	0,923
научн.	4,163	3,402	2,439	1,962	1,276	1,142	0,978	0,606	0,412	0,901

Таблица 4. Русский язык. XXI век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	4,121	3,307	2,367	1,483	0,843	0,656	0,452	0,345	0,182	0,956
офиц.-дел.	4,088	3,030	1,757	0,846	0,424	0,239	0,146	0,097	0,069	0,983
научн.	4,001	2,851	1,655	0,853	0,488	0,314	0,222	0,163	0,119	0,968

Значения, приведенные в таблицах, удобно представить в виде графиков, иллюстрирующих динамику изменения энтропии $H^{(m)}$ в зависимости от m для заданного стиля в различные временные интервалы.

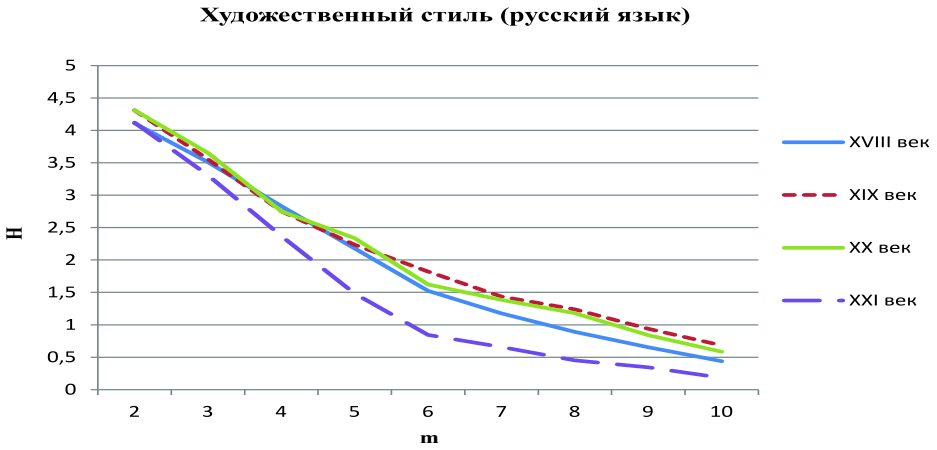

Рис. 1. Изменение энтропии $H^{(m)}$ художественных текстов в XVII–XXI вв

Следует отметить, что кривая изменения энтропии художественных текстов для русского языка не сильно выходит за области, соответствующие более ранним анализируемым временным интервалам, что может только свидетельствовать об относительной стабильности русской речи. С точки зрения теории информации художественный стиль с течением времени сохраняет свою информативность.

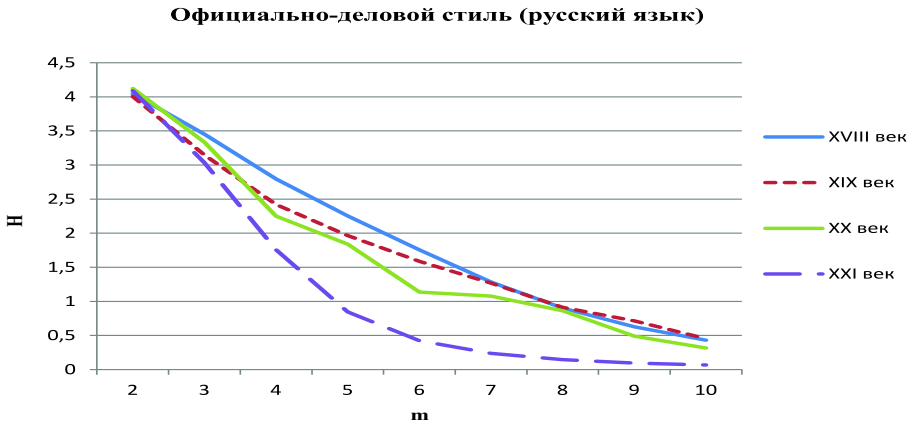

Рис. 2. Изменение энтропии $H^{(m)}$ текстов официально-делового стиля в XVIII–XXI вв

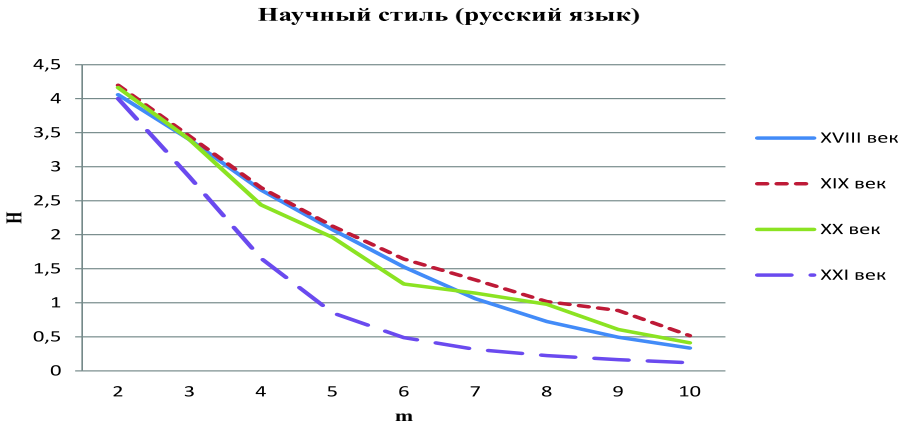

Рис. 3. Изменение энтропии научных текстов в XVIII–XXI вв

На рис. 2 ярко выражена динамика уменьшения энтропии в XXI веке. Этот факт, по всей видимости, объясняется тем, что с 1918 года в Советском Союзе были разработаны новые правила ведения служебной документации и введена единая форма бланков делового письма. В 20-е годы XX века была начата работа по созданию новых стандартов делового письма, что привело к появлению так называемых трафаретных текстов, которое и привело к уменьшению энтропии.

Далее приведем результаты вычислительных экспериментов, полученные для подобранных текстов, написанных на европейских языках: английском, немецком и французском.

Таблица 5. Английский язык. XVIII век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,916	3,255	2,696	2,108	1,740	1,306	1,139	0,763	0,534	0,863
офиц.-дел.	3,899	3,093	2,364	1,719	1,208	0,850	0,584	0,391	0,261	0,933
научн.	3,902	3,241	2,624	2,065	1,635	1,250	1,015	0,727	0,502	0,871

Таблица 6. Английский язык. XIX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,877	3,118	2,451	2,016	1,686	1,397	1,139	0,906	0,709	0,817
офиц.-дел.	3,839	3,157	2,362	1,733	1,242	0,888	0,619	0,415	0,283	0,926
научн.	3,852	3,090	2,417	1,982	1,564	1,272	1,029	0,736	0,511	0,867

Таблица 7. Английский язык. XX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,845	3,09	2,513	2,033	1,787	1,409	1,151	0,814	0,647	0,832
офиц.-дел.	3,779	2,811	2,022	1,674	1,471	0,969	0,752	0,631	0,445	0,882
научн.	3,806	2,986	2,386	1,837	1,525	1,146	0,971	0,705	0,478	0,874

Таблица 8. Английский язык. XXI век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,898	2,808	1,604	0,806	0,447	0,258	0,141	0,084	0,057	0,985
офиц.-дел.	3,692	2,263	0,988	0,453	0,270	0,148	0,081	0,063	0,038	0,990
научн.	3,819	2,524	1,109	0,554	0,348	0,161	0,092	0,070	0,053	0,986

Далее представлены графики изменения энтропии $H^{(m)}$ в зависимости от m, построенные по данным таблиц 5–8.

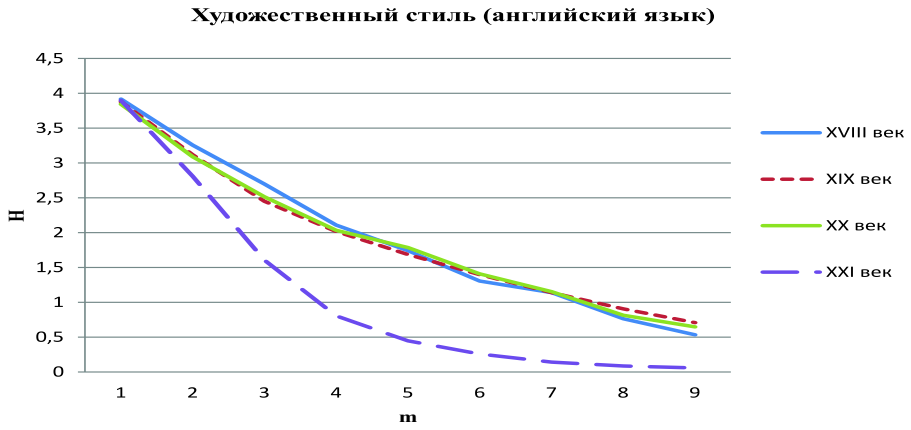

Рис. 4. Изменение энтропии $H^{(m)}$ художественных текстов в XVIII–XXI вв

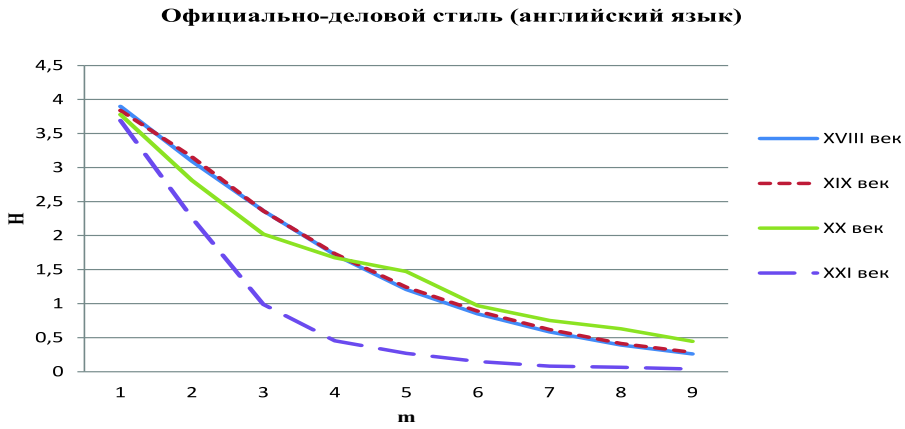

Рис. 5. Изменение энтропии $H^{(m)}$ текстов официально-делового стиля в XVIII–XXI вв

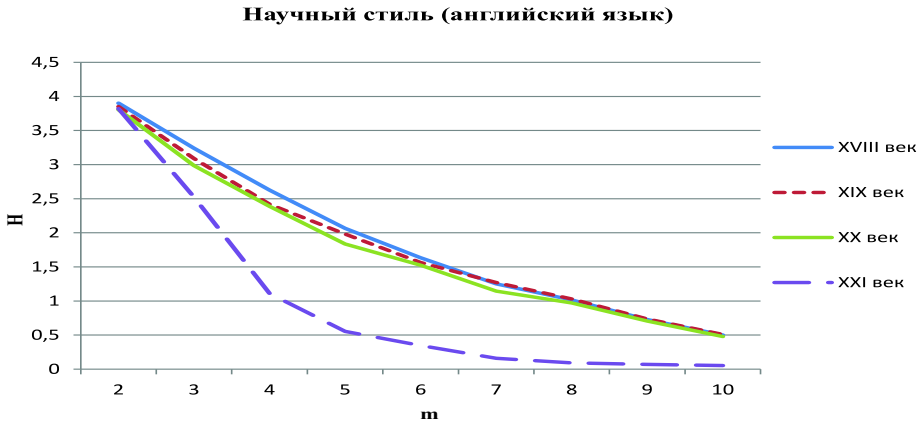

Рис. 6. Изменение энтропии $H^{(m)}$ научных текстов в XVIII–XXI вв

Приведем результаты экспериментов для немецкого языка.

Таблица 9. Немецкий язык. XVIII век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,829	3,129	2,558	2,067	1,885	1,428	1,334	1,113	0,967	0,747
офиц.-дел.	3,668	2,837	2,111	1,670	1,559	1,185	1,009	0,773	0,673	0,817
научн.	3,758	2,959	2,328	1,848	1,700	1,284	1,164	0,844	0,747	0,801

Таблица 10. Немецкий язык. XIX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,907	3,214	2,631	2,156	1,759	1,400	1,168	0,918	0,698	0,821
офиц.-дел.	3,723	2,902	2,212	1,682	1,294	0,983	0,699	0,593	0,443	0,881
научн.	3,774	3,003	2,387	1,824	1,425	1,125	0,898	0,773	0,586	0,845

Таблица 11. Немецкий язык. XX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,863	3,246	2,589	2,120	1,710	1,399	1,098	0,879	0,638	0,830
офиц.-дел.	3,594	2,867	2,114	1,645	1,312	0,961	0,648	0,580	0,407	0,822
научн.	3,745	3,068	2,302	1,817	1,446	1,089	0,848	0,720	0,557	0,851

Таблица 12. Немецкий язык. XXI век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,801	2,599	1,374	0,571	0,296	0,199	0,125	0,097	0,076	0,980
офиц.-дел.	3,621	2,174	0,930	0,370	0,169	0,125	0,074	0,057	0,030	0,992
научн.	3,728	2,360	1,135	0,490	0,204	0,164	0,097	0,068	0,048	0,987

Соответствующие графики, построенные по данным таблиц 9–12, имеют вид:

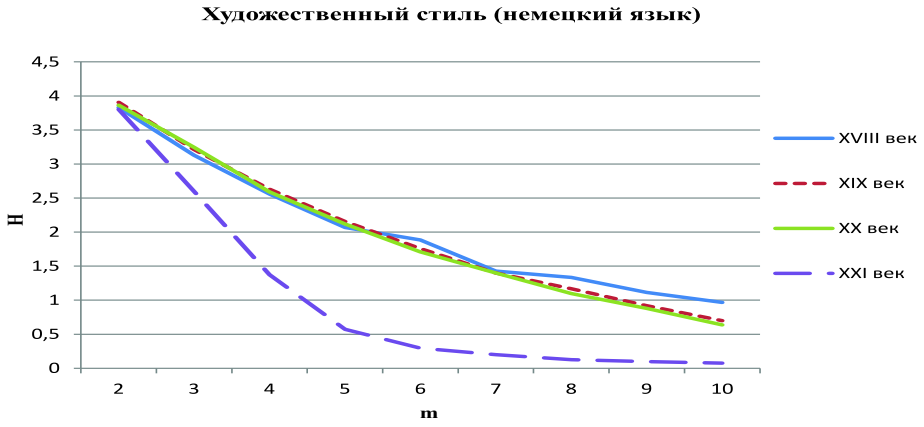

Рис. 7. Изменение энтропии $H^{(m)}$ художественных текстов в XVIII–XXI вв

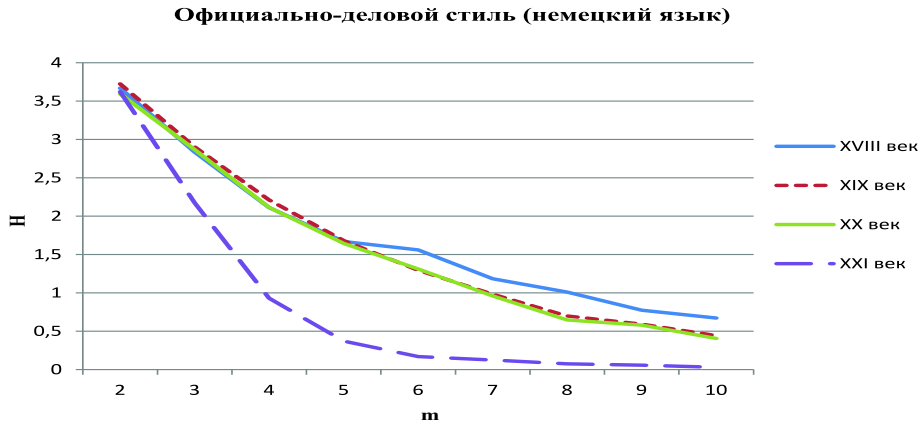

Рис. 8. Изменение энтропии $H^{(m)}$ текстов официально-делового стиля в XVIII–XXI вв

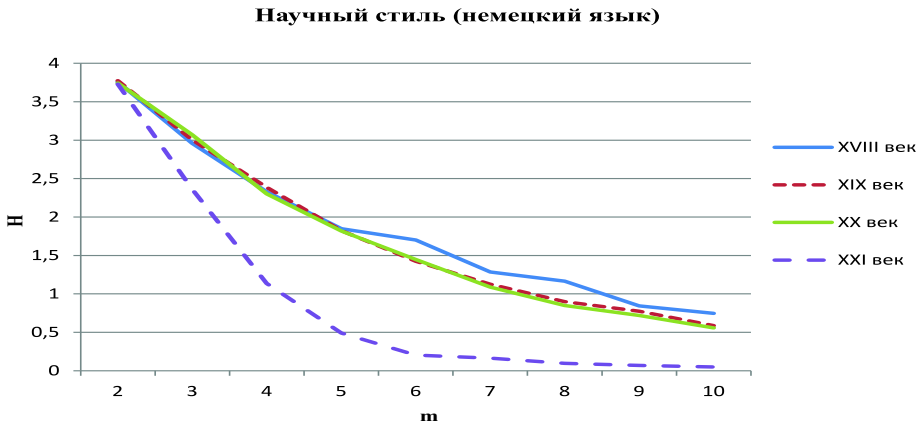

Рис. 9. Изменение энтропии $H^{(m)}$ научных текстов в XVIII–XXI вв

Приведем результаты для французского языка.

Таблица 13. Французский язык. XVIII

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,843	3,094	2,562	2,093	1,723	1,446	1,163	0,855	0,643	0,833
офиц.-дел.	3,789	2,984	2,428	2,037	1,670	1,267	0,854	0,569	0,367	0,903
научн.	3,803	3,044	2,497	2,059	1,711	1,377	1,063	0,742	0,535	0,859

Таблица 14. Французский язык. XIX

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,898	3,213	2,737	2,396	2,143	1,904	1,649	1,377	1,111	0,715
офиц.-дел.	3,760	3,013	2,423	1,976	1,646	1,345	1,054	0,802	0,554	0,862
научн.	3,796	3,191	2,656	2,209	1,830	1,472	1,117	0,797	0,605	0,841

Таблица 15. Французский язык. XX

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,894	3,226	2,716	2,241	1,813	1,428	1,134	0,860	0,636	0,837
офиц.-дел.	3,804	3,040	2,478	2,057	1,730	1,369	1,015	0,694	0,464	0,878
научн.	3,822	3,058	2,497	2,059	1,703	1,400	1,046	0,767	0,554	0,855

Таблица 16. Французский язык. XXI

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,724	2,838	1,741	1,439	0,782	0,508	0,378	0,260	0,148	0,960
офиц.-дел.	3,538	2,557	1,386	0,740	0,401	0,231	0,146	0,084	0,061	0,983
научн.	3,629	2,727	1,646	0,903	0,491	0,282	0,164	0,104	0,071	0,980

Соответствующие графики, построенные по данным таблиц 13–16, имеют вид:

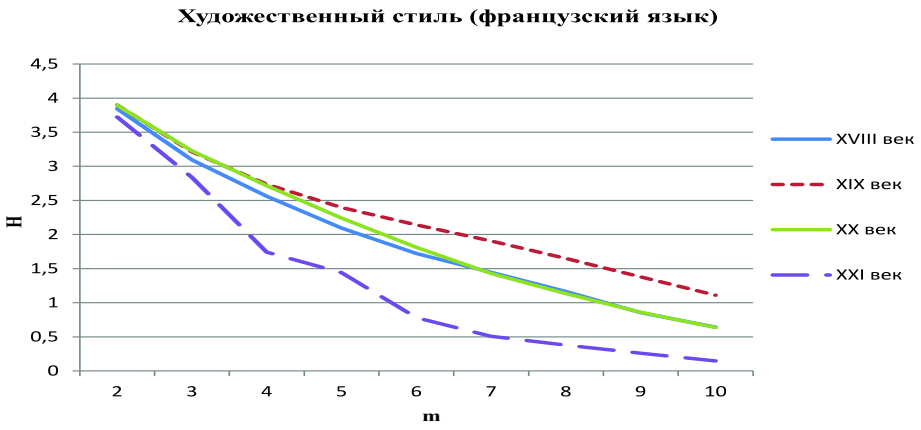

Рис. 10. Изменение энтропии $H^{(m)}$ художественных текстов в XVIII–XXI вв

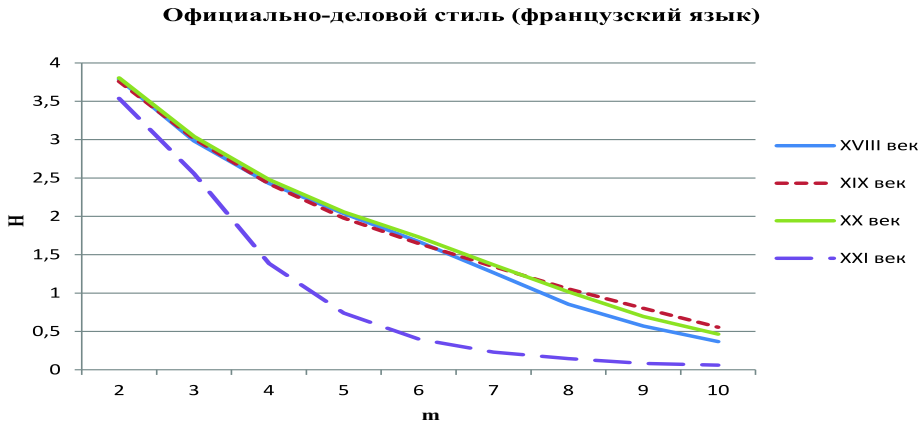

Рис. 11. Изменение энтропии $H^{(m)}$ текстов официально-делового стиля в XVIII–XXI вв

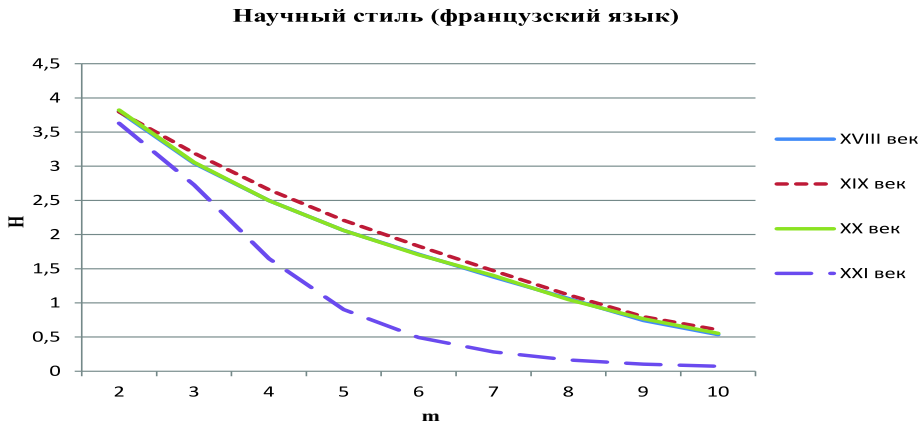

Рис. 12. Изменение энтропии $H^{(m)}$ научных текстов в XVIII–XXI вв

Из приведенных таблиц и графиков видна тенденция уменьшения энтропии выбранных представителей европейских языков во всех рассмотренных стилях за последний анализируемый период. Это может объясняться появлением ряда регламентирующих документов по оформлению документации, появлением различных социальных сетей, оказывающих необратимое негативное влияние на богатство языков, а также появлением общепринятого сленга в сети «Internet». Также в последнее время все большую популярность приобретают приложения коротких сообщений — например, таких, как Messenger или Twitter, что фактически приводит к упрощению современной лексики.

Рассмотрим теперь грузинский язык, не входящий в рассмотренные языковые группы. Напомним, что грузинский язык относится к картвельской группе кавказских языков, в которую входят мегрельский, сван-

ский, лазский языки. Грузинский язык является единственным среди иберийско-кавказских языков, который имеет древнюю письменность. Уникальное написание букв не сравнимо ни с одним алфавитом мира. Именно этими особенностями и объясняется научный интерес, проявленный авторами работы к грузинскому языку. Ниже приведены результаты соответствующих экспериментов.

Таблица 18. Грузинский язык. XVIII век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,980	3,240	2,540	1,970	1,528	1,167	0,889	0,668	0,488	0,877

Таблица 19. Грузинский язык. XIX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,974	3,308	2,661	2,152	1,728	1,344	1,01	0,756	0,542	0,864

Таблица 20. Грузинский язык. XX век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,914	3,222	2,53	1,987	1,521	1,133	0,887	0,651	0,471	0,880

Таблица 21. Грузинский язык. XXI век

стиль\m	2	3	4	5	6	7	8	9	10	R_{10}
худ.	3,552	1,890	0,694	0,247	0,137	0,074	0,063	0,050	0,041	0,988
офиц.-дел.	3,836	2,171	0,869	0,382	0,206	0,136	0,106	0,093	0,077	0,980

Соответствующий график изменения энтропии имеет вид:

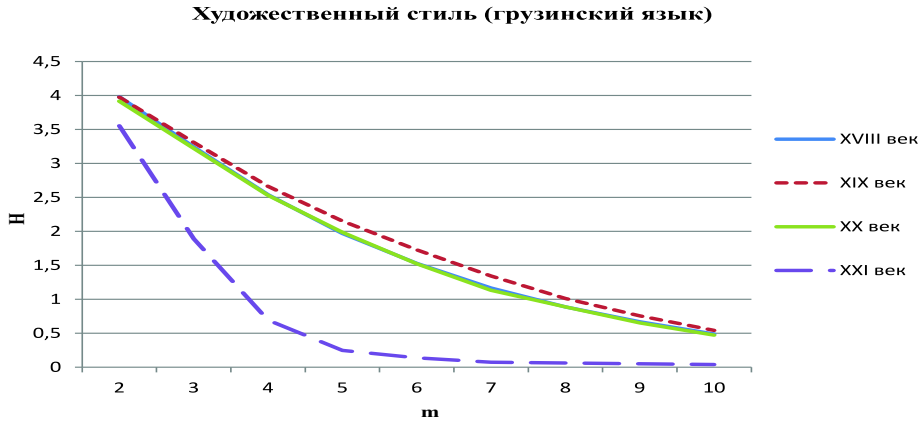

Рис. 13. Изменение энтропии $H^{(m)}$ художественных текстов в XVIII–XXI вв

Как видно из графика, в грузинском языке прослеживаются сходные тенденции уменьшения энтропии, как и в ранее проанализированных языках, что вызвано международной интеграцией как в политико-экономической, так и культурной сферах.

В свою очередь, авторы работы надеются, что, несмотря на развивающуюся динамику упрощения современной лексики и культуры, как русский, так и европейские языки, смогут выстоять, сохранив свою целостность. Полученные результаты позволяют в это верить.

СПИСОК ЛИТЕРАТУРЫ

1. *Духин А. А.* Теория информации. М.: Гелиос АРВ, 2007.
2. *Колесник Б. Д.*, *Полтырев Г. Ш.* Курс теории информации. М.: Наука, 1982.
3. *Пиотровский Р. Г.* Информационные измерения языка. Л.: Наука, 1968.
4. 4. *Савчук А. П.* Об оценках энтропии языка по Шеннону. — Теория вероятн. и ее примен., 1964, т. IX, в. 1, с. 154–157.
5. *Уотермен М. С.* Математические методы для анализа последовательностей ДНК. М.: Мир, 1999.
6. *Файнстейн А.* Основы теории информации. М.: ИЛ, 1960.
7. *Чечёта С. И.* Введение в дискретную теорию информации и кодирования. М.: МЦНМО, 2011.
8. *Шеннон К.* Работы по теории информации и кибернетике. М.: Издательство иностранной литературы, 1963.
9. *Buhlmann P.*, *Wyner A.* Variable length Markov chains. — Ann. Statist., 1999, v. 27, № 2, p. 480–513.
10. *Buhlmann P.*, *Machler M.* Variable length Markov chains: methodology, computing, and software. — J. Comput. Graph. Statist., 2004, v. 13, № 2, p. 435–455.
11. *Ching W. K.*, *Fung E. S.*, *Ng M. K.* High-order Markov chain models for categorical data sequences. — Naval Res. Logistics, 2004, v. 51, № 4, p. 557–574.
12. *Cover T. M.*, *Thomas J. A.* Elements of information theory. Wiley Interscience, 2006.
13. *Malyutov M.*, *Zhang Tong*, *Li Xin.*, *Li Yi.* Time series homogeneity tests via VLMC training. — Inform. Processes, 2013, v. 13, № 4, p. 201–214.
14. *Raftery A. E.* A model for High-Order Markov Chains. — J. Roy. Statist. Soc., 1985, v. B-47, № 3, p. 528–539.
15. *Rissanen J.*, *Harremoes P.*, *Forchhammer S.*, *Roos T.*, *Mullimake P.* Proceedings of The Eighth Workshop on Information Theoretic Methods in Science and Engineering. Ser. Publ. B. Report B-2015-1. Helsinki: Univ. Helsinki, Dept Comput. Sci., 2015.
16. *Yeung R. W.* A First Course in Information Theory. Dordrecht: Kluwer Academic/ Plenum Publishers, 2002.

Поступила в редакцию
25.XII.2015