УДК 004.43

С.Ш.Кажикенова, К.В.Мазиева, Е.Г.Шурыгина

Карагандинский государственный университет им. Е.А.Букетова (E-mail: sauleshka555@mail.ru)

Моделирование языка как сложной, динамичной, самоорганизующейся системы

В статье изложены исследования, обусловленные необходимостью изучения текстового материала различных жанров с целью его совершенствования. Предположено, что любой текст должен быть стилистически, грамматически, синтаксически оформлен грамотно, без лингвистических погрешностей. Авторами предложена идеальная лингвоматематическая модель для анализа структуры текста. Она построена на основе фундаментального закона сохранения суммы информации и энтропии с применением формулы Шеннона.

Ключевые слова: моделирование, самоорганизующаяся система, информация, энтропия, лингвосинергетика, формула Шеннона, формула Хартли, иерархическая структура, статистический метод, жанр текста, стиль текста, языковой уровень, модель.

Изучение языка методами теории информации стало перспективным научным направлением, исследующим сложные системы под углом зрения совершающихся в них процессов самоорганизации. В рамках этого направления происходит моделирование языка как сложной, динамичной, самоорганизующейся системы от неупорядоченного состояния к упорядоченному [1–4].

Подход к тексту как иерархической структуре позволяет рассматривать текст как с точки зрения анализа его составляющих, так и с точки зрения синтеза их на высшем языковом уровне.

Наши исследования обусловлены необходимостью изучения текстового материала различных жанров с целью его совершенствования. Любой текст должен быть стилистически, грамматически, синтаксически оформлен грамотно, без лингвистических погрешностей. Мы предлагаем идеальную лингвоматематическую модель для анализа структуры текста. Она построена на основе фундаментального закона сохранения суммы информации и энтропии с применением формулы Шеннона. При общей характеристике энтропийно-информационного анализа текстов (энтропия — мера беспорядка, а информация — мера снятия беспорядка) мы использовали статистическую формулу Шеннона для определения совершенства, гармонии текста:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i,$$

где p_i — вероятность обнаружения какой-либо единицы системы в их множестве N; $\sum_{i=1}^{N} p_i = 1,\ p_i \geq 0,$ $i = 1,2,...,N.$

До опубликования созданной Шенноном теории Хартли предложил определять количество максимальной энтропии по формуле

$$H_{\max} = \log_2 N.$$

Авторами проведен лингвистический анализ текста научного стиля речи [1], содержащего 500 знаков. Чтобы подсчитать p (вероятность) появления одной буквы в русском тексте, воспользуемся классической формулой определения вероятности. Для этого необходимо подсчитать, сколько букв содержится в этом тексте. Затем подсчитать, сколько раз встретилась отдельная буква в этом тексте. Тогда p (вероятность) появления одной буквы равна

$$P(a) = \frac{m}{n},$$

где n — число появления всех букв в этом тексте; m — число появления отдельной буквы.

Так как русский алфавит содержит 32 буквы (31 буква, 1 пробел), то максимальное значение энтропии текста, заключающегося в приеме одной буквы русского текста, при условии, что все буквы считаются одинаково вероятными, равно

$$Ho = \log 32 = 5 \text{ бит}.$$

Для более точного вычисления информации, содержащейся в одной букве русского текста научного стиля, надо знать вероятности появления различных букв в этом тексте. Для определения этих вероятностей рассмотрен отрывок из курса лекций по экономической теории [1], в которых основное внимание уделяется проблемам рыночного хозяйства. В лекциях наиболее детально рассматривается система экономических отношений и законов общественной жизни, различные типы рынков, различные точки зрения как отечественных, так и зарубежных ученых-экономистов. Выделенный отрывок «Рыночная инфраструктура» представляет собой текст научного стиля, в котором четко выражены признаки и особенности языка науки.

Итак, для вычисления информации научного текста были подсчитаны вероятности появления одной буквы, двухбуквенных, трехбуквенных, четырехбуквенных, пятибуквенных и шестибуквенных сочетаний в данном тексте. При подсчете учитывалась 31 буква русского алфавита (буквы *е* и *ё*, *ъ* и *ь* принимаются как одна буква) и пробел, все остальные знаки (скобки, кавычки, запятые и пр.) не рассматривались.

В ходе нашего исследования при подсчете числа повторений различных буквенных комбинаций в научном тексте на русском языке мы пришли к следующим показателям:

H_1	H_2	H_3	H_4	H_5	H_6
4, 364	2, 9766	0,78243	0,3426	0,0615	0,0537

Отсюда следует, что в научном тексте с увеличением информации происходит уменьшение степени неопределенности (энтропии). Однако на основе проведенных вычислений можно предположить, что в текстах научного стиля речи информации больше, чем в текстах других стилей.

Был проведен информационно-энтропийный анализ отрывка из газетной статьи «Бумага чиновнику ближе» Е.Ульянкиной из газеты «Новый Вестник».

Учитывая данные предыдущих расчетов, был проведен информационно-энтропийный анализ отрывка из Конституции РК, Основного Закона Казахстана, действующего со дня принятия Конституции Республики Казахстан на Всенародном референдуме 30 августа 1995 года.

Энтропия текста при учете одной буквы и двухбуквенных сочетаний в официально-деловом тексте меньше энтропии научного и публицистического текстов, энтропия текста при учете трехбуквенных сочетаний больше, энтропия при учете четырехбуквенных сочетаний меньше, энтропия при учете пятибуквенных и шестибуквенных немного превосходит количество энтропий научного и публицистического стилей. Проведенные расчеты подтверждают тот факт, что с увеличением информации происходит уменьшение степени неопределенности (энтропии) текста.

Таким образом, сопоставление данных показателей дает возможность заключить, что количество информации и энтропии в текстах официально-делового стиля принципиально отличается от текстов других стилей. Вероятно, этому способствуют отточенная форма изложения, регламентированность, стабильность и высокая информативность речевых средств.

Произведен информационно-энтропийный анализ текста разговорно-бытового стиля речи (письмо Н.В.Гоголя М.П.Погодину) [1], содержащего 500 знаков.

Для анализа текста стиля художественной литературы выбран рассказ А.П.Чехова «Крыжовник».

В результате были получены следующие значения (в битах):

H_1	H_2	H_3	H_4	H_5	H_6
4,2758	2,9952	1,2651	0,2986	0,0749	0,0252

Таким образом, результаты подсчета энтропии научного текста:

H_1	H_2	H_3	H_4	H_5	H_6
4, 364	2, 9766	0,78243	0,3426	0,0615	0,0537

Результаты подсчета энтропии публицистического текста:

H_1	H_2	H_3	H_4	H_5	H_6
4, 3742	3,0423	0,7895	0,5605	0,0451	0,0108

Результаты подсчета энтропии официально-делового текста:

H_1	H_2	H_3	H_4	H_5	H_6
4,2746	2,6721	0,9196	0,3290	0,1517	0,1046

Результаты подсчета энтропии разговорного текста:

H_1	H_2	H_3	H_4	H_5	H_6
4,3833	2,8383	1,0685	0,3683	0,0831	0,0630

Результаты подсчета энтропии художественного текста:

H_1	H_2	H_3	H_4	H_5	H_6
4,2758	2,9952	1,2651	0,2986	0,0749	0,0252

Как видим, энтропия текста при учете одной буквы и трехбуквенных сочетаний превышает показатели таких же сочетаний в четырех других стилях; энтропия текста при учете двухбуквенных сочетаний больше энтропии официально-делового текста и меньше энтропии научного, публицистического и художественных текстов; энтропия текста при учете четырехбуквенных сочетаний больше энтропии научного, художественного и официально-делового текста, но меньше энтропии текста публицистического стиля; энтропия текста при учете пятибуквенных и шестибуквенных сочетаний больше энтропии научного, художественного и публицистического стилей, но меньше энтропии официально-делового текста. Таким образом, сопоставление данных показателей дает возможность заключить, что количество информации и энтропии в текстах разговорного стиля принципиально отличается от текстов других стилей. Научный, художественный, официально-деловой и публицистический стили объединяет то, что они используются в сфере официального общения, а разговорный стиль используется в неофициальном, бытовом, повседневном общении.

Вероятно, этому способствуют неподготовленность и связанный с ней автоматизм построения высказывания; непринужденность и связанная с ней раскованность речи.

Результаты исследования заставляют полагать, что любой языковый текст, от единичного слова до объемного литературного произведения, может быть представлен как система, элементами которой являются отдельные буквы, а части представляют собой совокупность одинаковых букв. Соответственно, с помощью синергетической теории информации можно проводить структурный анализ произвольных текстов со стороны их хаотичности и упорядоченности по количеству и числу встречаемости отдельных букв.

Таким образом, всем рассмотренным материалом подчеркивается неразрывная взаимодополнительная связь детерминированной и вероятностной составляющих, из которых первая является доминирующей и обеспечивающей устойчивость, а вторая определяет наиболее тонкие изменения и оптимальную (а также максимальную) информационную емкость любых систем, в связи с чем вероятностно- детерминированный подход к их изучению представляется объективно необходимым.

Согласно закону сохранения суммы информации и энтропии, количество детерминированной информации текста рассчитывается как разность между максимально возможной энтропией и некоторым текущим значением энтропии.

Таким образом, информация и энтропия являются противоположно связанными характеристиками. А их взаимнодополнительное соотношение для любых самоорганизующихся систем диктуется законами сохранения суммы информации и энтропии и прогрессивного накопления информации при переходе с нижнего уровня организации на более высокий, вплоть до полной детерминации системы. В этом состоит основа сопоставления совершенства текстов, выраженной единой информационной характеристикой, со степенью детерминации идеальной иерархической системы на каждом уровне самоорганизации.

Предлагаемый информационно-энтропийный подход к определению объективной меры совершенства и полноты самоорганизации любых текстов можно рассматривать как развитие энтропийного анализа, при котором учитывалось только стремление энтропии к максимуму. В нашем подходе это стремление учитывается совместно с информационной составляющей, причем не в энергетических единицах, а в информационных битах.

Список литературы

1 *Кажикенова С.Ш., Оспанова Б.Р.* Информационно-энтропийный анализ структуры текста. — Караганда: Изд. КарГТУ, 2012. — 251 с.

2 *Кажикенова С.Ш., Оспанова Б.Р.* К вопросу о формировании концептуальной системы целевого языка в структуре коммуникативной компетенции // Язык и культура. — 2012. — № 3. — С. 111–121.

3 *Кажикенова С.Ш., Оспанова Б.Р.* О некоторых аспектах языковой модели в теории информации // Междунар. журн. экспериментального образования. — 2012. — № 8. — С. 115–120.

4 *Кажикенова С.Ш., Оспанова Б.Р.* Лингвосинергетический подход к исследованию текста как самоорганизующегося объекта // Хаос и структуры в нелинейных системах: Материалы междунар. науч.-практ. конф. (18–20 июня) / КарГУ. — Караганда, 2012. — С. 546–550.

С.Ш.Қажыкенова, К.В.Мазиева, Е.Г.Шурыгина

Тілді күрделі, динамикалық, өзін-өзі ұйымдастырушы жүйе ретінде үлгілеу

Мақалада зерттеуіміздің негізінде түрлі жанрлық мәтіндерді, оларды жетілдіру мақсатымен қарастырудың қажеттігі туралы айтылды. Кез келген мәтін стилистикалық, грамматикалық, синтаксистік жағынан, лингвистикалық ақаусыз дұрыс хатталған болуы керек. Мәтіннің құрылымын талдау үшін, біз идеалды лингвоматематикалық үлгіні ұсындық. Ол Шеннон формуласын қолдану арқылы, энтропия және ақпарат қосындысының сақталу заңы негізінде құрылған.

S.Sh.Kazhikenova, K.V.Maziyeva, Ye.G.Shurygina

The modeling of language like a complex dinamic self-organizing system

The researches due on need of studing of the text material of different genres with the purpose of its improvement are presented in this article. It is assumed that any text must be stulistically, grammatically, syntactically decorated correctly without linguistic errors. Offers an ideal lingua-mathematics model for the analysis of the text's structure. It is based on the fundamental law of conservation of the sum of information and entropy with using of Shennon's formula.

References

1 Kazhikenova S.Sh., Ospanova B.R. *Information and entropy analysis of structure of the text*, Karaganda: Publ. house of KarGTU, 2012, p. 251.

2 Kazhikenova S.Sh., Ospanova B.R. *To a question at about formation of conceptual system of target language in structure of communicative competence. Language and culture*, Tomsk, 2012, No. 3, p. 111–121.

3 Kazhikenova S.Sh., Ospanova B.R. *About some aspects of language model in information theory*. The International magazine of experimental education, 2012, No. 8, p 115–120.

4 Kazhikenova S.Sh., Ospanova B.R. *Lingvosinergetichesky approach to text research as self-organizing object*. Chaos and structures in nonlinear systems. Materials international scientific and practical conference (on June 18–20) / KarGU, Karaganda, 2012, p. 546–550.