Вестник Челябинского государственного университета. 2023. № 2 (472). С. 150–165.
ISSN 1994-2796 (print). ISSN 2782-4829 (online)
Bulletin of Chelyabinsk State University. 2023;(2(472):150-165. ISSN 1994-2796 (print). ISSN 2782-4829 (online)

Научная статья

УДК 80

СТОХАСТИЧНОСТЬ И ЭНТРОПИЯ В ЛИНГВИСТИКЕ

Елена Владимировна Шелестюк[1], Екатерина Алексеевна Щетинкина[2]✉

Челябинский государственный университет, Челябинск, Россия
[1]shelestiuk@yandex.ru, ORCID 0000-0003-4254-4439
[2]ekaterina8422@yandex.ru

Аннотация. Обсуждаются понятия стохастичности и энтропии текста, изучаются способы их измерения, рассматриваются спорные вопросы этих понятий и метрик. Выявляется связь стохастичности и энтропийности с категориями лингвистики текста (интегративностью и информативностью). Стохастичность текста — это непредсказуемость, хаотичность текстовых элементов, приводящая к эффекту новизны. Энтропия текста — это мера неопределенности содержания текста, она связана с объемом информации, содержащейся в данных: чем более неопределенны данные, тем больше информации требуется для их описания. Стохастичность описывает вероятностные характеристики данных, тогда как энтропия является мерой неопределенности, содержащейся в них. Поскольку энтропия и стохастичность имеют корреляцию, стохастичность может быть определена путем вычисления энтропии. Однако она также рассчитывается с использованием других методов и формул, в частности, перплексии, которая оценивает вероятность появления следующего слова в тексте на основе предыдущих слов. В лингвистике текста стохастичность и энтропия могут быть связаны с интегративностью (целостностью и связностью) текста. Они также могут определять информативность текста. Стохастичность может распространяться на текстовые информативные блоки и весь текст. Она создает семантическую сеть значений с внутренней целостностью и может определять семантическую и тематическую целостность текста (как часть интегративности). Уменьшение/рост энтропии связаны с уменьшением/усилением информации в тексте, т. о. энтропия имеет основополагающее значение для измерения информативности текста. Она также может измерять связность текста (как часть интегративности) на основе количества, глубины и повторяемости элементов *n*-грамм. Такой «эмпирико-синтаксический» подход измеряет связность и информативность по чисто формальным показателям. Однако энтропия и стохастичность не всегда точно отражают информативность и интегративность текста из-за факторов читабельности, понятности текста, его восприятия как истинного и разумного.

Ключевые слова: стохастичность текста, энтропия текста, перплексия, *n*-грамма, редукция энтропии, случайность, неопределенность, текстовые категории, интегративность, сюрпризал, информативность

Для цитирования: Шелестюк Е. В., Щетинкина Е. А. Стохастичность и энтропия в лингвистике // Вестник Челябинского государственного университета. 2023. № 2 (472). Филологические науки. Вып. 131. С. 150–165.

Original article

STOCHASTICITY AND ENTROPY IN LINGUISTICS

Elena V. Shelestyuk[1], Ekaterina A. Shchetinkina[2]✉

Chelyabinsk State University, Chelyabinsk, Russia
[1]shelestiuk@yandex.ru, ORCID 0000-0003-4254-4439
[2]ekaterina8422@yandex.ru

Abstract. The article discusses the concepts of text stochasticity and entropy of the text and methods for measuring these phenomena. They are two metrics providing a computational level explanation of how the probability and uncertainty of text units affect its cognitive processing. The connection of stochasticity and entropy with the categories of text linguistics — integrativity and informativeness — is revealed. Text stochasticity is unpredictability, randomness of text elements resulting in a novelty effect. Text entropy is a measure of the uncertainty associated with text data, it is related to the amount of information contained in the data; the more uncertain the data, the more information is required to describe it. Stochasticity describes the probabilistic characteristics of data, whereas entropy is a measure of the uncertainty contained in them. Since entropy and stochasticity have a correlation, stochasticity can

be determined through the calculation of entropy. However, it is also calculated using other methods and formulas, particularly, perplexity, which estimates the probability of the next word in a text based on the previous words. In text linguistics, stochasticity and entropy can be related to the integrativity (coherence and cohesion) of the text. They can also determine the informativeness of the text. Stochasticity can extend to text informative blocks and the entire text. It creates a semantic network of meanings with intrinsic integrity and can determine the semantic and thematic coherence of the text (as part of integrativity). A decrease/increase in entropy is associated with a decrease/increase in information in the text, i.e. entropy is fundamental for measuring the information content of the text. It can also measure text cohesion (as part of integrativity) based on the number, depth, and repeatability of n-gram elements. Such an "empirical-syntactic" approach measures integrativity and informativeness by purely formal indicators. However, entropy and stochasticity do not always accurately reflect the informativeness and integrativity of the text due to the factors of readability, comprehensibility of the text, its perception as true and rational.

Keywords: text stochasticity, text entropy, perplexity, n-gram, entropy reduction, randomness, uncertainty, text categories, integrativity, surprisal, informativeness

For citation: Shelestyuk EV, Schetinkina EA. Stochasticity and entropy in linguistics. *Bulletin of Chelyabinsk State University.* 2023;(2(472):150-165. (In Russ.).

Введение

Цель настоящей статьи заключается в рассмотрении понятий стохастичности и энтропии в приложении к лингвистике текста. Новизна работы определяется тем, что в настоящее время нет лингвистически релевантных сопоставлений этих смежных понятий теории информации и языкового моделирования. Гипотеза исследования заключается в том, что метрики стохастичности и энтропии можно эффективно использовать как для анализа свойств конкретного текста, так и для количественного и качественного анализа текстовых категорий. В свою очередь, накопленные данные могут позволить выявить закономерности текстов с разными метриками непредсказуемости/неожиданности/необычности/случайности/новизны и их восприятия в плане речевого воздействия. Методы исследования в данной статье включают анализ существующих на сегодняшний день теоретических данных об энтропии и стохастичности текста, в том числе рассмотрение противоречивых концепций в их отношении, приложение указанных понятий и метрик к текстовым категориям. Теоретическая значимость исследования определяется дальнейшей формализацией языковых (текстовых) данных с помощью моделирования и измерения информации. Практическая значимость исследования определяется применением в прикладной лингвистике, текстологии, создании и переводе текстов, обучении нейросетей и т. д.

Хаотичность и случайность как понятия, лежащие в основе энтропии и стохастичности

Разведем понятия хаотичности и стохастичности. Оба термина в общем смысле предполагают непредсказуемость, включая семантику необычности, случайности, новизны поведения элементов/информации. В математической статистике и информатике при обсуждении вопроса об отличии хаоса от стохастичности описываются разные характеристики систем.

Во-первых, хаотический процесс не обязательно случаен, но он может быть и случайным. Термин «хаотичный» используется для процессов, в которых практически невозможно точно описать долгосрочное поведение каким-либо значимым образом, зная начальное состояние (и даже зная некоторые правила перемещения между состояниями) в связи с большим и разнообразным влиянием на поведение элементов системы разных условий.

Что касается стохастики, то практически любой набор случайных величин, обычно индексируемых временем, может быть описан как «стохастический процесс». Но наибольший интерес представляют те стохастические процессы, в которых есть своего рода «клей», удерживающий случайные величины вместе, так что их долгосрочное поведение (с течением времени) может быть описано каким-то значимым образом.

Во-вторых, хаотический процесс проявляет чувствительность к начальным условиям. Случайный же процесс предполагает, что начальные условия функционирования системы могут не иметь большого значения [18].

Случайность может объясняться свойством независимых и одинаково распределенных случайных величин (IID, или НОР). IID был впервые использован в статистике, затем IID применяется в различных областях, таких как теория вероятностей, статистика, интеллектуальный анализ

данных и обработка сигналов. Правило IID гласит, что набор случайных величин независим и одинаково распределен, если каждая случайная величина имеет такое же распределение вероятностей, как и другие, и все они взаимно независимы [16]. В теории НОР последовательность случайных величин будет считаться стохастическим процессом. Несмотря на то, что ни одно наблюдение в последовательности не помогает предсказать другие, есть ряд интересных правил поведения случайных величин, описанных, например, законом больших чисел и центральной предельной теоремой[1].

Случайность может также описываться марковскими процессами. По определению Гаусса — Маркова, такие случайные элементы имеют своего рода одноступенчатую зависимость, когда вероятность каждого события в цепи зависит только от предыдущего события. Эта вероятность может стираться или не стираться с течением времени [45].

Хаотичность же, как упоминалось выше, описывает процессы, в которых, по существу, невозможно описать долгосрочное поведение системы каким-либо значимым образом, зная начальное состояние. Подбрасывание монеты считается «хаотичным» в том смысле, что невозможно предсказать, ляжет ли монета орлом или решкой, поскольку самые небольшие различия в скорости и вращении могут определить результат. Также прогнозы погоды более чем за три-четыре дня обычно бесполезны, поскольку существует слишком много мелких факторов, которые могут повлиять на то, будет ли дождь в конкретном месте в то или иное время [15].

Итак, хаотичность элементов не подчиняется правилам, не исчисляется (хотя энтропия — мера хаоса — может быть исчислена) и определяется множеством условий. Случайность же подчиняется правилам, исчисляется и не проявляет большой чувствительности к множественным условиям.

Говорить о стохастике мы можем однозначно, если на динамику процессов действуют факторы в рамках несинергических механизмов, направленных на достижение одного результата, единой цели управления. Примеров таких процессов много в неживой природе и технике. Мы можем говорить о жестких управляющих силах или о кооперации малых регуляторных систем, действующих в одном русле. Свободное падение с ускорением g на Земле в атмосфере будет описываться в рамках распределения Гаусса с модой, дисперсией и доверительным интервалом. Все законы физики, химии, ряд биозаконов протекают при доминировании одного или ряда синергетических законов, а другие процессы (движение воздуха, флуктуации плотности и т. д. в свободном падении, например) оказывают незначительное возмущение. Для таких процессов имеем выполнение стохастических закономерностей. Сейчас понятно, что если в системе действуют много законов (без доминантных), тем более если движущие силы процессов неоднородны и асинергичны, то говорить о стохастичности таких систем в их динамике нет смысла. Такие объекты есть в живой природе [1].

Случайность и хаотичность характерны и для языка и речи (текста). Однако хаотичность как таковая в языке и речи всё же редка, она переходит в измеряемую энтропию из-за ограниченного и постоянного количества условий (например, в речи — говорящие с учетом индивидуальных коммуникативных стилей, канал, сигнал, код, речевая (знаковая) ситуация, локуция) и достаточно выраженной прагматики, целеобусловленности. Поэтому и понятие энтропии (энтропийности) в естествознании и языкознании (шире — теории информации) несколько различается: если в первом случае энтропия определяется как мера хаоса, фиксирующая, но не способная однозначно описать и предсказать долгосрочное поведение объекта, то во втором случае энтропия случайных величин является мерой среднего уровня информации, неожиданности или неопределенности, присущего возможным результатам переменной, и описывается формулой К. Шеннона.

Понятие стохастичности. Стохастичность текста

Статистическая стохастичность и энтропия являются двумя важными концепциями в науке о данных и информации. Идеи стохастичности (stochasticity, surprisal, perplexity[2]) и энтро-

[1] = функциональная теорема о центральном пределе.

[2] Перплексия — термин, который был предложен Ф. Елинеком и Р. Мерсером [30]. Соответствующая метрика оценивает, насколько точно вероятностная модель может предсказать образцы текста, и измеряет уровень неопределенности в языковой модели. Это понятие можно рассматривать как обратную вероятность текста в предположении определенной языковой модели или как обратную вероятность тестовой выборки, которая нормализуется по числу слов. С другой стороны, сюрпризал К. Шеннона измеряет количество информации, которую мы

пии (entropy) были предложены независимо друг от друга в качестве релевантных (и альтернативных) гипотез, связывающих вероятностные языковые модели и когнитивные усилия по обработке сложностей текста [12; 13; 14; 19–23; 39; 44; 52]. Эти метрики обеспечивают объяснение на вычислительном уровне того, как вероятность и неопределенность единиц текста влияют на его когнитивную обработку.

Стохастичность относится к вероятностной природе данных. В статистической терминологии стохастический процесс — это процесс, который может быть описан вероятностными методами, то есть методами, которые определяют вероятность различных событий в рамках этого процесса. Стохастичность текста определяется как непредсказуемость, неожиданность, необычность, случайность появления элементов в тексте и возникающий в результате этого эффект новизны. Однако, по существу, уже само понятие стохастичности является диалектичным: в разных контекстах оно может означать ожидаемое (вероятное, известное и т. п.) и неожиданное (невероятное, неизвестное и т. п.), а также само понятие ожидания. Обусловлено это тем, что в среднестатистическом тексте, то есть в тексте с обычной понятностью, условно каждый пятый элемент является неожиданным, новым, а четыре элемента справа и слева от него — ожидаемыми и известными.

Энтропия — это мера неопределенности, связанной с данными: чем более неопределенны данные, тем выше энтропия. Точнее, энтропия связана с количеством информации, содержащейся в данных, то есть чем более неопределенны данные, тем больше информации требуется, чтобы описать их. Например, в случае, когда все возможные значения равновероятны, то есть неопределенность максимальна, количество информации равно количеству возможных значений, умноженному на количество бит, необходимых для представления каждого значения.

Таким образом, основная разница между стохастичностью и энтропией заключается в том, что стохастичность описывает вероятностные характеристики данных, в то время как энтропия представляет собой меру неопределенности, содержащейся в этих данных. Это взаимосвязанные, но разные величины.

Если говорить непосредственно о тексте, то и здесь присутствует корреляция: чем ниже стохастичность элементов текста, тем ниже и энтропийность самого текста. Стохастичность текста характеризует, насколько случайным образом расположены буквы/символы/слова в тексте. Если текст имеет высокую стохастичность, то буквы/символы/слова будут встречаться случайным образом, и у него будет более высокая энтропия. С другой стороны, если стохастичность элементов текста ниже, то текст будет иметь более определенную структуру, и он будет иметь более низкую энтропию. Впрочем, такая корреляция не является абсолютной закономерностью, и каждый текст может иметь свою специфику. Также следует учитывать, что стохастичность и энтропия — это математические понятия, которые не всегда могут быть легко применимы к естественному языку и текстам, которые написаны на нем.

Существует программное обеспечение, которое способно рассчитать стохастичность текста. Например, в языке программирования Python есть библиотека nltk, которая может быть использована для вычисления стохастичности текста. В этой библиотеке есть функция nltk.probability.FreqDist, которая позволяет определить частоту встречаемости каждого слова в тексте, а затем на основе этой информации можно рассчитать стохастичность. Также есть программы Textalyser, TextSTAT, Voyant Tools и другие, которые могут выполнять анализ текста и рассчитывать стохастичность. Эти программы могут быть особенно полезны для исследования больших объемов текста или для выполнения анализа научных статей.

При создании модели языка для расчета стохастичности текста используются различные варианты обучающих данных, такие как словари, корпусы, статистические данные и т. д. Они позволяют на основе анализа большого количества текстов определить частоту использования конкретных слов и их сочетания, что помогает улучшить точность расчета стохастичности текста. Некоторые модели языка могут быть сформированы заранее, чтобы использоваться в виде

получаем, получая конкретное сообщение. Он вычисляется с помощью логарифма обратной вероятности события и сообщает нам, насколько неожиданным или удивительным является это событие. Перплексия и сюрпризал тесно связаны между собой, но имеют различные формальные определения и используются для разных целей. Применение перплексии связано с задачами языкового моделирования и распознавания речи, в то время как сюрпризал используется в теории информации и смежных областях. В данной работе, однако, мы будем рассматривать эти термины как синонимы.

библиотек или API для расчета стохастичности текста, в то время как другие модели могут быть созданы индивидуально под конкретные задачи и нужды пользователей.

Итак, в основном энтропия и стохастичность имеют достаточно четкую корреляцию. Исходя из этого, стохастичность может быть определена через вычисление энтропии. Для нас существенно, однако, что можно вычислить стохастичность текста без вычисления его энтропии, используя другие методы и формулы. Одним из них может быть вычисление перплексии (perplexion, surprisal), которая также является статистической мерой оценки стохастичности текста. Перплексия оценивает вероятность следующего слова в тексте на основе предыдущих слов. То есть, чем выше перплексия, тем меньше предсказуемости в тексте, тем менее связанными являются слова в тексте.

Перплексия как мера того, насколько хорошо языковая модель предсказывает последовательность слов, вычисляется как обратная вероятность тестового набора, нормализуемая по количеству слов. Перплексия тестовой последовательности $w_1, ..., w_N$ в модели θ представлена так:

$$PPL = \sqrt[N]{\frac{1}{\mathrm{P}(w_1,..., w_N \mid \theta)}} = \sqrt[N]{\frac{1}{\prod_{i=1}^{N}\mathrm{P}(w_i \mid w_1,..., w_{i-1}, \theta)}}$$

или в экспоненциальной форме: $PPL = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 \mathrm{P}(w_i|w_1,..., w_{i-1},\theta)}$ [1]. Концепция перплексии существует уже давно, но конкретная формула для ее вычисления в контексте обработки естественного языка часто приписывается Ф. Елинеку и Р. Мерсеру (1980), которые предложили использовать перплексию в качестве оценочной метрики для языковых моделей в статье “Interpolated estimation of Markov source parameters from sparse data” [30].

Перплексия также вычисляется автоматически на основе значений частотности слов, полученных с помощью nltk.probability.FreqDist, и оценивает, насколько «удивительным» может быть следующее слово, учитывая предыдущие слова в тексте. В качестве меры стохастичности текста меньшее значение перплексии указывает на более предсказуемый или менее «случайный» текст, а большее значение перплексии указывает на более стохастический или более «случайный» текст. Чтобы вычислить перплексию в Python, можно использовать функцию perplexity() из класса nltk.LanguageModel. Эта функция принимает текст и модель языка, построенную на основе корпуса текста и объекта класса nltk.probability.FreqDist и возвращает значение перплексии. Функция nltk.probability.FreqDist определяет частотность каждого слова в тексте, но не может вычислить стохастичность текста напрямую. Однако на основе значений частотности слов, полученных с помощью nltk.probability.FreqDist, можно вычислить некоторые статистические показатели, такие как энтропия или перплексия, которые используются для оценки стохастичности текста. Чтобы выполнить этот расчет, необходимо использовать дополнительные функции из библиотеки nltk или другие математические формулы. Также для лингвистического анализа используются различные метрики, связанные со стохастичностью, например, Bigram Frequency или TF-IDF (Term Frequency-Inverse Document Frequency), которые способны измерять взаимосвязь между словами и частоту появления слов в тексте.

Можно создать также программу по вычислению стохастичности и по формуле А. Н. Мороховского $G = H/D$, где G — коэффициент стохастичности, H — количество недетерминированных элементов в тексте, а D — детерминированных. При этом известно, что средний коэффициент стохастичности равен 1/4, то есть 0,25 [7]. Для этого нужно написать функцию, которая могла бы рассчитать количество недетерминированных и детерминированных элементов в тексте, а затем использовать формулу $G = H/D$ для определения коэффициента стохастичности. Чтобы сделать программу более точной, можно использовать текстовые модели для сопоставления значимых пяти элементов текста. Например, можно использовать модели языка n-грамм, которые учитывают частоту комбинаций из n элементов (в данном случае $n = 5$), чтобы более точно определять, какие элементы следует считать недетерминированными. Средний коэффициент стохастичности, равный 0,25, может быть использован как опорное значение для сравнения полученного результата. Если коэффициент стохастичности выше 0,25, то текст считается более стохастическим, а если ниже — менее стохастическим.

Стохастичность и перплексия могут оцениваться на уровне отдельных слов, но также могут быть использованы для оценки стохастичности текста в целом. При этом их оценка может учи-

[1] Кудинов М. С. Статистическое моделирование русского языка с помощью нейронных сетей : дис. ... канд. техн. наук : 05.13.17; [Место защиты: Федер. исслед. центр Информатика и упр.]. М., 2016. 20 с.

тывать языковую модель, построенную на основе более крупных единиц текста, таких как фразы, предложения или даже тематические поля. Точнее, перплексия может использоваться для оценки стохастичности текста на различных уровнях, включая уровень отдельных слов и уровень тематических полей. Однако при оценке перплексии на уровне тематических полей или фраз необходимо использовать соответствующую языковую модель, построенную на более крупных единицах текста.

Энтропия в информатике; энтропия текста и способы ее измерения

Информационная энтропия — мера неопределенности некоторой системы, в частности непредсказуемость появления какой-либо единицы первичного алфавита (словаря, реестра). Несколько затемняя картину, К. Шеннон понимал энтропию как информационную избыточность [40], в то же время утверждая, что прирост информации равен утраченной неопределенности. А. Н. Тырсин и др. цитирует Джона фон Неймана («Никто не знает, чем на самом деле является энтропия») и пишет, что «несмотря на частое употребление этого термина, использование энтропии для моделирования сложных систем, в отличие от термодинамики, не формализовано и носит качественный характер». Вместе с тем «энтропия может выступать в роли универсального параметра и идеально подходит для решения рассматриваемых задач о поведении сложных стохастических систем» [8; 47].

Есть известные формулы расчета энтропии. В классической статье А. Н. Колмогорова «Три подхода к определению понятия количества информации» (1965) рассматриваются три способа это сделать: 1) комбинаторный (информация по Р. Хартли), 2) вероятностный (энтропия К. Шеннона), 3) алгоритмический (колмогоровская сложность). Опишем две первые формулы.

Формула Хартли описывает систему с равновероятностной встречаемостью элементов. Последовательность N величин с K-значными символами (в зависимости от числа букв алфавита) может принимать $V = K^N$ значений. Если последовательность принимает V равновероятных значений, то для их записи требуется не менее чем $N = \log_K V$ символов. Поэтому для случайной величины, принимающей конечное число значений, естественно определить «собственную информацию» (меру определенности) как логарифм от обратной вероятности $I_i = \log_K(1/P_i) = -\log_K(P_i)$, где i — номер значения (события), K — основание алфавита (списка). Чем вероятнее событие, тем меньше информации оно приносит: $\log_2(1) = 0$, а, например, $\log_2(3) = 1$, …

Вероятностный подход, предложенный Шенноном в 1948 г., обобщает определение Хартли на случай, когда не все элементы множества являются равнозначными. Энтропией Шеннона называют средневзвешенное количество информации в сообщениях источника. Современная математическая статистика выделяет «энтропию информации» и «энтропию взаимосвязи». Оба понятия используются и для статистического исследования текста. Рассмотрим информационную энтропию как меру неопределенности системы S. Приводится формула $H(S) = -\sum_{i=1}^{m} p_i \log p_i$, где p_1, …, p_m — вероятности того, что система S принимает конечное число соответствующих значений x_1, …, x_m, так $\sum_{i=1}^{m} p_i = 1,\ p_i \geq 0, i = 1, 2, ..., m$, что от основания логарифма в (1) зависит единица измерения информации — бит, трит и др. [6; 8; 47].

Энтропия взаимосвязи двух случайных величин означает меру взаимного хаоса или взаимной упорядоченности рассматриваемых случайных величин. Для того, чтобы идея использования энтропии взаимосвязи для анализа связи текстов на естественном языке могла быть использована, требуется привести рассматриваемые тексты в вид, к которому метод расчета энтропии взаимосвязи может быть применен, что фактически требует преобразования текста на естественном языке к виду численной случайной величины, или, другими словами, вектора случайных значений [47]. В нашем исследовании акцент ставится на энтропии информации, то есть событий, но не взаимосвязей.

В лингвистике текста энтропийность информации в настоящее время успешно исследуется с помощью отечественных программ — LanA-Key, инструмента сбора статистических моделей SMAT (Statistical Model Acquisition Tool), которые выдают список 1-, 2-, 3- и 4-словных именных фраз с показателем их релевантности в обработанном тексте.

Остановимся на программе SMAT. На основе формулы К. Шеннона была создана программа вычисления энтропийности текста. N-граммы в SMAT — это ряды слов (не синтагмы[1]), двоичные n-граммы 1–2; 2–3; 3–4 и т. д.; троичные

[1] Чтобы получить из текста синтагмы, требуются иные экстракторы энтропии.

123, 234, 345, 456, 567… и так далее. Вычисление энтропии с помощью SMAT — чистая эвристика для каждого текста. Программа может быть использована для разных целей, в том числе для автоматической атрибуции текста.

Если энтропийность по существу представляет собой непредсказуемость, стремление к хаосу, то ее противоположностью в лингвистике является клишированность. Под клишированностью понимается повторяемость элементов текста, в частности, повторяемость слов и синтагм у автора. Чем больше энтропийность, тем выше степень неопределенности текста, чем больше клишированность, тем меньше степень неопределенности текста.

Имеются маркированные клише — идиомы, прецедентные высказывания, афоризмы, и немаркированные клише, то есть фактическое повторение слов и синтагм у автора. Маркированные клише таким образом сводятся к фразеологическим или стереотипным речевым явлениям, а немаркированные — к повторам, полным, частичным или тематическим у автора. К клише могут относиться: 1) крылатые выражения, 2) авторские повторы внутри текста, 3) однооформленные лексические единицы в тексте или авторское однообразие. О клишированности может свидетельствовать использование частотных, широкозначных или «всеобщих» слов и синтаксические особенности текста. Можно вывести формулу клишированности, например, количество маркированных клише плюс количество немаркированных клише, деленное на число слов.

При определении текстовой энтропии важно также понятие релевантности — количественной характеристики, которая позволяет сортировать извлеченные единицы в зависимости от ряда их свойств. Показатель релевантности зависит от частоты самой фразы, совокупной частоты ее компонентов, длины, частоты вхождения в более длинные фразы, места в тексте (больший вес присваивается фразам, встречающимся в начале текста), количества параграфов, в которых встретилась фраза, а также общепринятых критериев статистической релевантности [10; 41; 43]. Именные фразы, к которым относятся и однословные существительные, — самый частотный слой лексики, наиболее тесно связанный с содержанием текста.

Между компонентами *n*-грамм энтропия минимальна, между самими же *n*-граммами энтропия может достигать высоких значений. Энтропия и клишированность определяются повторяемостью элементов *n*-грамм. Наиболее функциональным является понятие редукции энтропии, которая определяет понимание смысла высказывания/текста.

Диалектика понятий и метрик стохастичности и энтропии (энтропийности) текста: обзор литературы

Подведем итог предыдущих параграфов. Основная разница между стохастичностью и энтропией заключается в том, что стохастичность описывает вероятностные характеристики данных, в то время как энтропия представляет собой меру неопределенности, содержащейся в этих данных. Это взаимосвязанные, но разные величины. Исходя из того, что в основном энтропия и стохастичность имеют достаточно четкую корреляцию, стохастичность может быть определена через вычисление энтропии и наоборот. Для нас существенно, однако, что эти величины можно вычислять и сравнивать автономно. Так, можно вычислить стохастичность текста без вычисления его энтропии. Одним из них может быть вычисление перплексии (perplexion, surprisal), которая оценивает вероятность следующего слова в тексте на основе предыдущих слов и вычисляется по формуле $PP(W) = 2^{-\Sigma \log_2 p(w_i)}$, где $PP(W)$ — перплексивность тестового множества W, а $p(w_i)$ — вероятность, присвоенная каждому слову w_i в тестовом наборе языковой модели (формула Ф. Елинека и Р. Мерсера). Бесплатным программным инструментом, который может измерить перплексию текста, является функция перплексии в наборе средств естественного языка (NLTK) в Python. Библиотеки языкового моделирования, такие как Stanford CoreNLP и Apache OpenNLP, также предоставляют функциональные возможности для вычисления перплексии. Кроме того, некоторые фреймворки машинного обучения, такие как TensorFlow и PyTorch, также имеют встроенные функции для вычисления перплексии. Энтропия текста может быть определена как среднее количество битов или информационных единиц, необходимых для представления каждого слова в тексте. Формула энтропии текста определяется как среднее количество информационных единиц (бит или шанс) для представления каждого слова в тексте. Она вычисляется по формуле: $H = -\Sigma P(w) \log_2 P(w)$, где H — энтропия, $P(w)$ — вероятность каждого слова в тексте (формула К. Шеннона). Энтропия текста также может быть измерена с помощью программных приложений. Подобно вычислению перплексии, для

вычисления энтропии текста можно использовать библиотеки языкового моделирования, такие как NLTK, Stanford CoreNLP и OpenNLP, фреймворки машинного обучения, такие как TensorFlow и PyTorch. Энтропийность текста также измеряется с помощью программ LanA-Key и SMAT.

Формула перплексии текста и формула энтропии текста имеют сходство, но не являются понятиями эквивалентными. Обе формулы используются для измерения количества информации, содержащейся в тексте, и оба термина относятся к теории информации. Однако формула перплексии является мерой того, насколько сложно понять текст, тогда как энтропия текста является мерой количества информации, содержащейся в тексте.

Формула перплексии связана с понятием вероятности, что следующее слово в тексте будет таким, каким ожидается. Чем ниже вероятность, тем более непонятен текст. Для вычисления перплексии нужно знать вероятность тестового набора (записана под корнем в знаменателе), то есть принимается во внимание фактор семантики. Формула энтропии, с другой стороны, измеряет разнообразие слов и степень их использования в тексте.

Таким образом, можно сказать, что формула перплексии и формула энтропии не выводятся одна из другой, но обе они используются для оценки того, насколько информативен и понятен текст.

Стохастичность и энтропия (редукция энтропии) являются важными явлениями, определяющими простоту/сложность обработки информации текстов и лежащими в основе вероятностных языковых моделей. Ниже приведем рассуждения исследователей о диалектике понятия стохастичности и понятия энтропии (редукции энтропии) в понимании смысла текста.

Язык обрабатывается более или менее пословно, при этом некоторые слова вызывают больше усилий по обработке смысла, чем другие, что отражается в более длительном времени чтения (reading time, RT). С когнитивной точки зрения, происходит разворачивающийся во времени процесс пословного обновления интерпретации [51]. Существующие в информатике модели обработки текста, как правило, основываются на лингвистическом опыте и ограничиваются его учетом. Однако сложность обработки (=когнитивные усилия), вызванная каждым словом, зависит не только от предыдущего лингвистического опыта, но и от общих знаний о мире. Большинство моделей обработки информации не могут объяснить влияние знания о мире в процессе понимания.

Многие современные модели ориентируются на человеческое понимание, на смысловое восприятие текста. Авторы [51] предлагают дифференциальные метрики обработки текста: оценку стохастичности и редукции энтропии. Обе метрики измеряются, исходя из лингвистического опыта и знания о мире понимающего субъекта. Интерпретация понимается как производная от обработки единиц «ожиданий и неожиданного» [ibid.]. Показано, как когнитивно-ориентированная модель редукции энтропии проявляет себя во время интерпретации и понимания текста и как редукция энтропии и стохастичность отражаются в различных поведенческих эффектах [20; 21].

Утверждается, что стохастичность и энтропийная редукция происходят из одного и того же когнитивного процесса, но отражают различные аспекты этого процесса. Стохастичность показывает ожидание одного состояния за другим и восприятие его как ожидаемого/неожиданного, а редукция энтропии подтверждает конечное состояние процесса и тем самым свидетельствует об обнаружении смысла, о логичности вывода. Состояние в случае текста — это то же, что событие в общей теории стохастики и энтропии, типично под событием/состоянием понимается появление заданного члена предложения (=слова) в цепи из пяти слов.

Язык обрабатывается более или менее пословно. Под влиянием теории коммуникации было высказано предположение, что информативность слова прямо пропорциональна усилиям по его обработке, которые оно вызывает. Одним из способов количественной оценки информативности слова является использование понятия сюрпризала, которое является метрикой, количественно определяющей ожидаемость слова [26–28; 33]. При этом чем менее ожидаемо слово в данном контексте, тем выше его «самоинформация» (self-information) [50].

Много экспериментальных результатов показывают, что когнитивная обработка сложности отдельных слов зависит не только от их вероятности как части (локального) лингвистического контекста, но и от более широкого дискурса и визуального контекста, а также от общих знаний о мире [11; 14; 17; 24; 28; 29; 31; 32; 34–37; 48; 49 и др.]. Следовательно, чтобы объяснить эти результаты с точки зрения информативности слов, теоретико-информационные метрики стохастичности и сокращения энтропии должны учитывать вероятностную структуру мира, выходящую

за рамки только вербального контекста, то есть лингвистических сигналов. Они должны быть либо дополнены вероятностным понятием экстралингвистического знания, либо пересмотрены с точки зрения лежащих в их основе когнитивных процессов.

В [50; 51] представлена модель понимания языка, в которой оценки стохастичности выводятся из вероятностных распределенных репрезентаций значений (смысловых представлений), цепочки из которых выстраиваются на пословной основе (так называемый «процесс инкрементального языкового понимания»). Эти репрезентации, основанные на структуре распределенного ситуационного пространства (Distributed Situation-state Space framework) [22; 23], используются для создания экземпляров ситуационных моделей, которые позволяют делать выводы, основанные на знаниях о мире. Таким образом, языковая модель расширяется за счет когнитивно-ориентированной модели понимания, в рамках которой человек или нейронные сети оперируют как языковым опытом (историей лингвистической информации), так и знанием о мире (вероятностным знанием, отраженным в представлениях) [48; 49]. Показано, что процесс понимания «чувствителен» к обоим этим источникам информации, предлагается объяснять то, как лингвистический опыт и знание мира влияют на обработку сложностей текста, с помощью репрезентативного и алгоритмического правила Марра [34].

Ряд источников полагают, что редукция энтропии является релевантным и при этом независимым от стохастичности предиктором при обработке когнитивных сложностей текста. Типично языковая энтропия определяется анализом лингвистических структур, при этом используются такие методы, как вероятностная контекстно-свободная грамматика состояний [26–28], частей речи [39] или отдельных слов [19–23]. Вместе с тем энтропию можно рассматривать как величину неопределенности относительно положения дел в мире. То есть уменьшение энтропии при понимании, например, слова w_t количественно показывает, сколько неопределенности в отношении текущего положения дел устраняется обработкой слова w_t. Такой подход эмпирически подтверждается исследованием ситуативного понимания языка, при котором при предъявлении реципиентам вербально-визуального контента менялся лишь визуальный контекст, а вербальный текст оставался неизменным, тем самым сохранялась лингвистическая константа стохастичности [46]. Эксперимент показал, что неподходящий экстралингвистический визуальный контекст повышает референтную энтропию по отношению к лингвистическому контексту, что приводит к увеличению усилий по обработке идентичных высказываний.

Стохастичность количественно определяет, насколько вероятна следующая точка, заданная предыдущей и, таким образом, насколько ожидаемым был ввод. Таким образом, стохастичность можно рассматривать как отражение ожидания от состояния к состоянию, где входные данные перемещают модель в ожидаемые или неожиданные точки в пространстве. Энтропия, в свою очередь, количественно определяет, насколько вероятно каждое полностью определенное положение дел, составляющее пространство значений, учитывая текущую точку в пространстве. Таким образом, редукция энтропии фактически является метрикой подтверждения конечного состояния, где более высокое уменьшение неопределенности в отношении предложений, то есть более сильное подтверждение сообщаемого состояния дел, приводит к более высокому уменьшению энтропии. Эта характеристика, по-видимому, соответствует последним теориям понимания текста, в которой понятие валидации — процесса оценки согласованности поступающей лингвистической информации с предыдущим лингвистическим контекстом и общими знаниями о мире — играет центральную роль [17; 37; 38; 42].

Когнитивно-ориентированная вычислительная модель понимания [50; 51] показывает, как метрики стохастичности (неожиданности) и энтропии связаны с процессом понимания и какие дифференциальные прогнозы эти метрики делают относительно понимания. Стохастичность и редукция энтропии отражает один и тот же когнитивный процесс — процесс понимания как своеобразную навигацию через пространство значений. Однако они отражают различные аспекты этого процесса: ожидание некого состояния за предшествующим состоянием (связанное со стохастичностью) vs подтверждение конечного состояния (связанное с уменьшением энтропии). Таким образом, ожидаемое/неожиданное (стохастичность) и сокращение энтропии в процессе понимания дифференциально отражают эффекты лингвистического опыта и знания о мире, которые интегрируются на уровне интерпретации и понимания смысла информации. Существенно, что стохастичность и редукция энтропии могут лежать в основе нейронной сети (simple recurrent neural network), которая выстра-

ивает репрезентации значений (смысловые представления) в виде словосочетаний и высказываний на инкрементной пословной основе. Эксперимент с подобной обучаемой сетью и продемонстрирован в [ibid.].

Вышесказанное согласуется с парадигмой понимания языка, в которой неотъемлемой частью инкрементной пословной обработки текстовой информации является прагматический вывод. Фактически, можно утверждать, что когнитивно-ориентированная модель воплощает подход, в котором понимание завершается прагматическим выводом. Буквальное пропозициональное содержание высказывания не имеет особого статуса, важны только вероятностные выводы, которые вытекают из обработки *всего* высказывания (формально складывающегося из буквального пропозиционального содержания). Таким образом, в понимании информатиков, в частности Noortje J. Venhuizen et al. [ibid.], стохастичность, которая вытекает непосредственно из двух последующих точек в пространстве значений, эффективно отражает локальное изменение вероятности выведенных предложений, поскольку она учитывает только выводы, содержащиеся в этих точках. Редукция энтропии, в свою очередь, отражает вероятностность предполагаемых предложений на глобальном уровне, то есть по отношению к полному набору возможных выводов о фрагментах реальности, которые могут быть сделаны. Редукция энтропии рассматривает степень необычности (энтропийность) между этими точками, которая влияет на вероятность всех возможных выводов.

Связь стохастичности и энтропийности с категориями лингвистики текста (интегративностью и информативностью)

Учитывая приведенные выше точки зрения, изложим соображения по поводу приложения стохастичности и энтропийности к лингвистике текста. Оба явления соотносятся с необычным/обыденным, то есть с оправданием или неоправданием ожиданий воспринимающего. Однако стохастичность не может быть сведена лишь к семантической согласованности/несогласованности близлежащих единиц или блоков информации, но может распространяться на всё высказывание или текст. Это свойство, или категория, текста создает семантическую сеть значений, обладающую внутренней целостностью, и в качестве такого семантического и тематического единства она отражает фрагмент реального либо воображаемого мира. Стохастичность действительно «чувствительна» к лингвистическому опыту, но не в меньшей степени она чувствительна к знаниям о действительности. Существенно, что стохастичность нам представляется в большей степени семантическим, нежели синтаксическим механизмом внутритекстовой связи, она определяет семантическую/тематическую целостность (когерентность), а не синтаксическую связность (когезию).

На уровне компьютерных программ при существующих в настоящее время наработках, стохастичность, на наш взгляд, оптимально может исчисляться двумя методами: 1) в русле рассмотрения текста как системы с равновероятностной встречаемостью элементов по Р. Хартли, отбирается каждый 5-й (в идеале — и 2-й, 3-й, …, 7-й) значимый элемент текста, а затем определяется степень семантического/тематического сходства (близости) этих элементов; 2) в русле рассмотрения текста как системы со средневзвешенным количеством информации по К. Шеннону производится семантический (SEO) или контент-анализ; затем весь массив ранжированной лексики делится на сегменты ядра, ближней периферии и дальней периферии семантического пространства; следующим шагом является, как и в первом способе, определение степени семантического/тематического сходства (близости) этих элементов. Если первые этапы предложенных методов достаточно легко произвести с помощью автоматизированных программ, то определение семантического сходства (близости) элементов и выделение тематических полей в массиве лексики до сих пор достаточно трудноразрешимы.

Что касается энтропии, то в расширительном плане редукция энтропии может означать сокращение неопределенности и понимание прагматического смысла высказывания/текста. Однако вполне целесообразно, на наш взгляд, и синтаксическое понимание энтропии как величины, зависящей от числа, глубины и повторяемости элементов *n*-грамм. «Синтаксический» подход позволяет предсказать энтропийность текста: чем больше коротких *n*-грамм (в пределе, соответствующие одному слову), тем выше энтропия, и наоборот: чем больше в тексте длинных *n*-грамм, тем ниже энтропия.

В лингвистике текста стохастичность и энтропию можно сопрягать со связностью текста, или его когезией, и целостностью (цельностью) текста, или его когерентностью. Когезия показывает значимые синтаксические (точнее, синтактические)

отношения внутри текста, то есть касается синтаксиса (синтактики), когерентность же отражает значимые лексико-семантические отношения внутри текста — синонимию, метонимию, гипогиперонимию, принадлежность к одним семантическим (тематическим) полям слов, используемых в тексте, то есть касается лексики (семантики).

Когерентность и когезия, в свою очередь, являются подкатегориями такой категории текста, как интегративность. Рассмотрим формальные средства выражения интегративности текста, то есть средства выражения его когезии, или связности. Связный текст, отмечает А. А. Леонтьев, воспринимается читателем как некоторое единство на основании разного рода формальных признаков — средств соотнесения частей текста между собой. Эти признаки отличаются ретроспективным характером, так как они не задаются автором заранее, а появляются по мере создания и восприятия текста [4]. В то же время внутритекстовое соотнесение может осуществляться в двух разных направлениях: анафорически — при соотнесении с предыдущим фрагментом текста и катафорически — при соотнесении с последующим фрагментом текста. Кроме того, внутритекстовое соотнесение осуществляется двумя путями: непосредственно, как прямое указание на конкретную часть текста, и опосредствованно — через соотнесение содержания различных фрагментов текста. Непосредственное соотнесение частей текста между собой характерно преимущественно для текстов жесткого и узуального типа, в частности для научной прозы, для публицистического же и художественного текста нередки имплицитные, подтекстовые связи и связки в виде модальных операторов (даже, всё равно, считается и проч.). К связующим средствам универсального характера относятся: 1) элементы текста, указывающие на смысловую неполноту, или синсемантию, его фрагментов; 2) различного рода повторы; 3) темарематическое соотнесение предложений и абзацев в дискурсе; 4) стилистические приемы разных уровней; 5) коммуникативная соотнесенность компонентов текста; 6) композиционно-структурные особенности текста в целом [2].

Рассмотрим содержательно-смысловые средства выражения интегративности текста, то есть средства выражения его когерентности, или цельности (целостности). Когерентность текста носит (лексико)-семантический характер и предполагает прежде всего смысловое единство текста. Целостный (цельный) текст, по мнению А. А. Леонтьева, можно определить как текст, который при переходе от одной последовательной ступени компрессии к другой, более глубокой, каждый раз сохраняет смысловое тождество, то есть инвариантное значение, информативное ядро текста, лишаясь лишь периферийных, второстепенных элементов [4]. Наиболее абстрагированное выражение информативного ядра получило название макроструктуры текста, которая вычленяется в результате применения к линейной смысловой структуре текста серии операций (трансформаций) свертывания и служит наиболее кратким выражением содержания текста [3; 8]. В лингвистике текста отмечается, что содержательная целостность текста создается взаимодействием следующих факторов: 1) наличием коммуникативной интенции автора; 2) тематическим единством текста; 3) объединяющей функцией «образа автора»; 4) связующей ролью различных типов выдвижения в тексте; 5) объединяющей функцией выразительных средств и стилистических приемов, реализующихся одновременно в пределах единицы текста и всего текста в целом; 6) композиционно-жанровым единством [2].

Итак, связность и целостность — важные категории текста, которые в конечном счете определяются соотношением закономерных, предсказуемых и случайных, непредсказуемых связей и элементов в тексте, то есть такими его характеристиками, как энтропия и стохастичность. Формализация упомянутых категорий в функциях энтропии и стохастичности могла бы продвинуть исследования в междисциплинарной области текстовой информатики. Например, с точки зрения формального исчисления связности как синтаксической предсказуемости/случайности, перспективным объектом являются элементы текста, указывающие на смысловую неполноту его фрагментов, синтаксически связанные фразы, относительно законченные фразеологические сочетания. С точки зрения формального исчисления целостности как семантической предсказуемости/случайности, перспективным объектом является тематическое (и семантическое) единство лексики текста в виде тематических (семантических) полей и т. п.

Очевидно, что стохастика и энтропия текста могут определять и информативность текста. Информативность текста есть «степень его смыслосодержательной новизны для читателя, которая заключена в теме и авторской концепции, систе-

ма авторских оценок предмета мысли» [1]. Чем ниже связность и целостность текста и чем выше определяющие их энтропийность и стохастичность, тем ниже предсказуемость, выше новизна и тем, возможно, информативнее текст. Таким образом, связь между информативностью, с одной стороны, и связностью и целостностью текста, с другой, в определенных случаях обратно пропорциональна.

Метриками информативности могут быть как коэффициент стохастичности или перплексии текста, так и энтропия (редукция энтропии), индуцированные словами. Эти метрики показывают, насколько непредсказуем текст: чем больше разнообразие слов в тексте, тем выше стохастичность и энтропия; если текст содержит много повторяющихся слов и фраз, то стохастичность и энтропия будут низкими. Метрики стохастичности и энтропии могут быть использованы, например, для автоматической оценки качества машинного перевода, где более информативный перевод должен иметь более высокую стохастичность и энтропию. Однако эти метрики не всегда точно отражают качество текста, поскольку высокая стохастичность и энтропия могут также означать низкую читабельность и понятность текста.

Также следует учитывать, помимо факторов целостности и связности текста, факторы истинности и рациональности (включающие семантику логичности и предметной отнесенности), тесно связанные с информативностью. Без их учета связь энтропии и стохастичности с информативностью оказывается неоднозначной.

Заключение

Понятия и метрики стохастичности и энтропии могут быть полезными в лингвистике текста для анализа и моделирования языка. Эти метрики могут использоваться для измерения неопределенности и разнообразия языкового материала в тексте. Например, энтропия Шеннона может быть использована для измерения средней неопределенности символов (например, букв или слов) в тексте. Тексты с более высокой энтропией обычно более разнообразны по своим символам, в то время как тексты с более низкой энтропией будут иметь более предсказуемые и повторяющиеся символы. Другая метрика, которая может быть полезной в лингвистическом анализе — это перплексия; она может помочь определить, насколько хорошо модель подходит для конкретного текста. В целом использование метрик стохастичности и энтропии может помочь лингвистам лучше понять структуру и свойства текста. Это может быть полезно в том числе для анализа категорий текста, таких как интегративность и информативность, текстовой грамматики и семантики, определения авторства текста, анализа стиля и прочих характеристик текста.

Список источников

1. Адайкин В. И. и др. Новый метод идентификации хаотических и стохастических параметров экосреды // Вестник новых медицинских технологий. 2006. Т. XIII, № 2. С. 39–41.

2. Бабайлова А. Э. Текст как продукт, средство и объект коммуникации при обучении неродному языку: социопсихолингвистические аспекты / под ред. А. А. Леонтьева. Саратов : Изд-во Сарат. ун-та, 1987.

3. Воробьева О. П. Стилистика текста // Стилистика английского языка : учебник для студентов инов и фак. иностр. яз. / А. Н. Мороховский, О. П. Воробьева, Н. И. Лихошерст, З. В. Тимошенко. Киев : Выща шк., 1991. С. 201–235.

4. Жинкин Н. И. Язык — речь — творчество : исслед. по семиотике, психолингвистике, поэтике: (избр. тр.). М. : Лабиринт, 1998. 364 с.

5. Леонтьев А. А. Основы психолингвистики. М. : Смысл, 1997. 221 с.

6. Марченко А. Д., Тырсин А. Н. Использование энтропии взаимосвязи в анализе текстов на естественном языке // Современные наукоемкие технологии. 2021. № 6-1. С. 67–73.

7. Мороховский А. Н., Воробьева О. П., Лихошерст Н. И., Тимошенко З. В. Стилистика английского языка. Киев : Вища школа, 1984. 247 с.

8. Тырсин А. Н. Энтропийное моделирование многомерных стохастических систем. Воронеж : Научная книга, 2016. 156 с.

9. Шахнарович А. М. Общая психолингвистика : учебник пособие. М. : Изд-во РОУ, 1995. 96 с.

10. Шереметьева С. О. Об использовании программ обработки текста для обучения иностранным языкам // Вестник ЮУрГУ. 2012. № 25. Серия «Лингвистика», вып. 15. С. 56–59.

11. Штернберг М. И. Синергетика и биология // Вопросы философии. 1999. № 2. С. 95–108.

12. Blache P., Rauzy S. Predicting linguistic difficulty by means of a morpho-syntactic probabilistic model // Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC-2011), Singapore, 16–18 December 2011. P. 160–167.

13. Boston M. F., Hale J. T., Kliegl R., Patil U., Vasishth S. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus // J. Eye Mov. Res. 2008. № 2. P. 1–12.

14. Brouwer H., Fitz H., Hoeks J. Modeling the Noun Phrase versus Sentence Coordination Ambiguity in Dutch: Evidence from Surprisal Theory // Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics; Association for Computational Linguistics : Uppsala, Sweden, 2010. P. 72–80.

15. BruceET (https://math.stackexchange.com/users/221800/bruceet), Difference between stochastic process and chaotic system, URL: https://math.stackexchange.com/q/1349805

16. Clauset A. A brief primer on probability distributions. Santa Fe Institute, 2011.

17. Cook A. E., Myers J. L. Processing discourse roles in scripted narratives: The influences of context and world knowledge // J. Mem. Lang. 2004. № 50. P. 268–288.

18. Farimani Foad S. What is the difference between chaotic systems and stochastic systems? URL: https://www.quora.com/What-is-the-difference-between-chaotic-systems-and-stochastic-systems/answer/Foad-S-Farimani?ch=2&srid=iETG

19. Frank S. L. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing // Proceedings of the 31st Annual Conference of the Cognitive Science Society; Cognitive Science Society: Austin, TX, USA, 2009. P. 1139–1144.

20. Frank S. L. Uncertainty reduction as a measure of cognitive load in sentence comprehension // Cogn. Sci. 2013. № 5. P. 475–494.

21. Frank S. L. Uncertainty reduction as a measure of cognitive processing effort // Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics; Association for Computational Linguistics. Stroudsburg, PA, USA, 2010. P. 81–89.

22. Frank S. L., Haselager W. F., van Rooij I. Connectionist semantic systematicity // Cognition. 2009. № 110. P. 358–379.

23. Frank S. L., Koppen M., Noordman L. G., Vonk W. Modeling knowledge-based inferences in story comprehension // Cogn. Sci. 2003. № 27. P. 875–910.

24. Garrod S., Terras M. The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution // J. Mem. Lang. 2000. № 42. P. 526–544.

25. Hale J. T. A probabilistic Earley parser as a psycholinguistic model // Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies; Association for Computational Linguistics. Stroudsburg, PA, USA, 2001. P.1–8.

26. Hale J. T. The information conveyed by words in sentences // Psycholinguist. Res. 2003. № 32. P. 101–123.

27. Hale J. T. Uncertainty about the rest of the sentence // Cogn. Sci. 2006. № 30. P. 643–672.

28. Hale J. T. What a rational parser would do // Cogn. Sci. 2011. № 35. P. 399–443.

29. Hess D. J., Foss D. J., Carroll P. Effects of global and local context on lexical processing during language comprehension. // Exp. Psychol. Gen. 1995. № 124. P. 62–82.

30. Jelinek F., Mercer R. Interpolated estimation of Markov source parameters from sparse data // Proc. of the Workshop on Pattern Recognition in Practice. Amsterdam, 1980. P. 381–397.

31. Knoeferle P., Crocker M. W., Scheepers C., Pickering M. J. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events // Cognition. 2005. № 95. P. 95–127.

32. Knoeferle P., Habets B., Crocker M. W., Münte T. F. Visual scenes trigger immediate syntactic reanalysis: Evidence from ERPs during situated spoken comprehension // Cereb. Cortex. 2008. № 18. P. 789–795.

33. Levy R. Expectation-based syntactic comprehension // Cognition. 2008. № 106. P. 1126–1177.

34. Marr D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman : San Francisco, CA, USA. 1982.

35. Morris R. K. Lexical and message-level sentence context effects on fixation times in reading // Exp. Psychol. Learn. Mem. Cogn. 1994. № 20. P. 92–102.

36. Myers J. L., O'Brien E. J. Accessing the discourse representation during reading // Discourse Process. 1998. № 26. P. 131–157.

37. O'Brien E. J., Cook A. E. Coherence threshold and the continuity of processing: The RI–Val model of comprehension // Discourse Process. 2016. № 53. P. 326–338.

38. Richter T. Validation and comprehension of text information: Two sides of the same coin // Discourse Process. 2015. № 52. P. 337–355.

39. Roark B., Bachrach A., Cardenas C., Pallier C. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 1. Association for Computational Linguistics : Stroudsburg, PA, USA, 2009. Pp. 324–333.

40. Shannon C. E. A mathematical theory of communication // Bell Syst. Tech. J. 1948. № 27. P. 379–423.

41. Sheremetyeva S. On Extracting Multiword NP Terminology for MT // Proceedings of the 13th Conference of European Association for Machine Translation. Barcelona, Spain. P. 205–212.

42. Singer, M. Validation in reading comprehension // Curr. Dir. Psychol. Sci. 2013. № 22. P. 361–366.

43. Smadja F. Retrieving collocations from text. Xtract // Computational Linguistics. 1993. № 7 (4). P. 143–177.

44. Smith N. J., Levy R. Optimal Processing Times in Reading: A Formal Model and Empirical Investigation // Proceedings of the 30th Annual Meeting of the Cognitive Science Society. Cognitive Science Society : Austin, TX, USA, 2008. P. 595–600.

45. Stock J. H., Watson M. W. Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals // Introduction to Econometrics. 3. Addison-Wesley, 2011. P. 163–164.

46. Tourtouri E. N., Delogu F., Sikos L., Crocker M. W. Rational over-specification in visually-situated comprehension and production // J. Cult. Cogn. Sci. 2019. doi:10.1007/s41809-019-00032-6.

47. Tyrsin A. N., Sokolova I. S. Entropy-probabilistic modeling of Gaussian stochastic systems // Matem. Mod. 2012. Vol. 24. Number 1. P. 88–102.

48. van Berkum J. J. A., Brown C. M., Zwitserlood P., Kooijman V., Hagoort P. Anticipating upcoming words in discourse : Evidence from ERPs and reading times // J. Exp. Psychol. Learn. Mem. Cogn. 2005. № 31. P. 443–467.

49. van Berkum J. J. A., Zwitserlood P., Hagoort P., Brown C. M. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect // Cogn. Brain Res. 2003. № 17. P. 701–718.

50. Venhuizen Noortje J., Crocker Matthew W., Brouwer Harm. Semantic Entropy in Language Comprehension // Entropy. 2019. № 21 (12). P. 1159.

51. Venhuizen N. J., Crocker M. W., Brouwer H. Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience // Discourse Process. 2019. № 56. P. 229–255. doi:10.1080/0163853X.2018.1448677.

52. Wu S., Bachrach A., Cardenas C., Schuler W. Complexity metrics in an incremental right-corner parser // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010. P. 1189–1198.

References

1. Adajkin VI. et al. Novyj metod identifikacii haoticheskih i stohasticheskih parametrov ekosredy. *Vestnik novyh medicinskih tekhnologij*, 2006;XIII(2):39-41. (In Russ.).

2. Babajlova AE. Tekst kak produkt, sredstvo i ob'ekt kommunikacii pri obuchenii nerodnomu yazyku: sociopsiholingvisticheskie aspekty. Saratov; 1987. (In Russ.).

3. Vorob'eva OP. Stilistika teksta. In: Morohovskij AN, Vorob'eva OP, Lihosherst NI, Timoshenko ZV. Stilistika anglijskogo yazyka: uchebnik. Kiev; 1991. Pp. 201–235. (In Russ.).

4. Zhinkin NI. Yazyk — rech' — tvorchestvo: issledovanija po semiotike, psiholingvistike, poetike. Moscow, Labirint; 1998. 364 p. (In Russ.).

5. Leont'ev AA. Osnovy psiholingvistiki. Moscow, Smysl; 1997. 221 p. (In Russ.).

6. Marchenko AD, Tyrsin AN. Ispol'zovanie entropii vzaimosvyazi v analize tekstov na estestvennom yazyke. *Sovremennye naukoemkie tekhnologii*. 2021;(6-1):67-73. (In Russ.).

7. Morohovskij AN, Vorob'eva OP, Lihosherst NI, Timoshenko ZV. Stilistika anglijskogo yazyka. Kiev, Vishcha shkola; 1984. 247 p. (In Russ.).

8. Tyrsin AN. Entropijnoe modelirovanie mnogomernyh stohasticheskih sistem. Voronezh, Nauchnaya kniga; 2016. 156 p. (In Russ.).

9. Shahnarovich AM. Obshchaya psiholingvistika: uchebnoye posobie. Moscow; 1995. 96 p. (In Russ.).

10. Sheremet'eva SO. Ob ispol'zovanii programm obrabotki teksta dlya obucheniya inostrannym yazykam. *Vestnik YUUrGU*. 2012;(15):56-59. (In Russ.).

11. Shternberg MI. Sinergetika i biologiya. *Voprosy filosofii*. 1999;(2):95-108. (In Russ.).

12. Blache P, Rauzy S. Predicting linguistic difficulty by means of a morpho-syntactic probabilistic model. In: Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC-2011), Singapore, 16–18 December; 2011. Pp. 160–167.

13. Boston MF, Hale JT, Kliegl R, Patil U, Vasishth S. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *J. Eye Mov. Res*. 2008;(2):1-12.

14. Brouwer H, Fitz H, Hoeks J. Modeling the Noun Phrase versus Sentence Coordination Ambiguity in Dutch: Evidence from Surprisal Theory. In: Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics; Association for Computational Linguistics, Uppsala, Sweden; 2010. Pp. 72–80.

15. BruceET (https://math.stackexchange.com/users/221800/bruceet), Difference between stochastic process and chaotic system. Available from: https://math.stackexchange.com/q/1349805

16. Clauset A. A brief primer on probability distributions. Santa Fe Institute; 2011.

17. Cook AE, Myers JL. Processing discourse roles in scripted narratives: The influences of context and world knowledge. *J. Mem. Lang*. 2004;(50):268-288.

18. Farimani Foad S. What is the difference between chaotic systems and stochastic systems? Available from: https://www.quora.com/What-is-the-difference-between-chaotic-systems-and-stochastic-systems/answer/Foad-S-Farimani?ch=2&srid=iETG

19. Frank SL. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society; Cognitive Science Society: Austin, TX, USA; 2009. Pp. 1139–1144.

20. Frank SL. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Cogn. Sci*. 2013;(5):475-494.

21. Frank SL. Uncertainty reduction as a measure of cognitive processing effort. In: Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics; Association for Computational Linguistics. Stroudsburg, PA, USA; 2010. Pp. 81–89.

22. Frank SL, Haselager WF, van Rooij I. Connectionist semantic systematicity. *Cognition*. 2009;(110):358-379.

23. Frank SL, Koppen M, Noordman LG, Vonk W. Modeling knowledge-based inferences in story comprehension. *Cogn. Sci*. 2003;(27):875-910.

24. Garrod S, Terras M. The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. *J. Mem. Lang*. 2000;(42):526-544.

25. Hale JT. A probabilistic Earley parser as a psycholinguistic model. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies; Association for Computational Linguistics. Stroudsburg, PA, USA; 2001. Pp. 1–8.

26. Hale JT. The information conveyed by words in sentences. *Psycholinguist. Res*. 2003;(32):101-123.

27. Hale JT. Uncertainty about the rest of the sentence. *Cogn. Sci*. 2006;(30):643-672.

28. Hale JT. What a rational parser would do. *Cogn. Sci*. 2011;(35):399-443.

29. Hess DJ, Foss DJ, Carroll P. Effects of global and local context on lexical processing during language comprehension. *Exp. Psychol. Gen.* 1995;(124);62-82.

30. Jelinek F, Mercer R. Interpolated estimation of Markov source parameters from sparse data. In: Proc. of the Workshop on Pattern Recognition in Practice. Amsterdam; 1980. Pp. 381–397.

31. Knoeferle P, Crocker MW, Scheepers C, Pickering MJ. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*. 2005;(95):95-127.

32. Knoeferle P, Habets B, Crocker MW, Münte TF. Visual scenes trigger immediate syntactic reanalysis: Evidence from ERPs during situated spoken comprehension. *Cereb. Cortex*. 2008;(18):789-795.

33. Levy R. Expectation-based syntactic comprehension. *Cognition*. 2008;(106):1126-1177.

34. Marr D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman, San Francisco, CA, USA; 1982.

35. Morris RK. Lexical and message-level sentence context effects on fixation times in reading. *Exp. Psychol. Learn. Mem. Cogn*. 1994;(20):92-102.

36. Myers JL, O'Brien EJ. Accessing the discourse representation during reading. *Discourse Process*. 1998;(26):131-157.

37. O'Brien EJ, Cook AE. Coherence threshold and the continuity of processing: The RI–Val model of comprehension. *Discourse Process*. 2016;(53):326-338.

38. Richter T. Validation and comprehension of text information: Two sides of the same coin. *Discourse Process.* 2015;(52):337-355.

39. Roark B, Bachrach A, Cardenas C, Pallier C. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 1. Association for Computational Linguistics. Stroudsburg, PA, USA; 2009. Pp. 324–333.

40. Shannon CE. A mathematical theory of communication. *Bell Syst. Tech. J.* 1948;(27):379-423.

41. Sheremetyeva S. On Extracting Multiword NP Terminology for MT. In: Proceedings of the 13th Conference of European Association for Machine Translation. Barcelona, Spain. Pp. 205–212.

42. Singer M. Validation in reading comprehension. *Curr. Dir. Psychol. Sci*. 2013;(22):361-366.

43. Smadja F. Retrieving collocations from text. Xtract. *Computational Linguistics*. 1993;(7(4)):143-177.

44. Smith NJ, Levy R. Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. In: Proceedings of the 30th Annual Meeting of the Cognitive Science Society. Cognitive Science Society : Austin, TX, USA; 2008. Pp. 595–600.

45. Stock JH, Watson MW. Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals // Introduction to Econometrics. 3. Addison-Wesley; 2011. Pp. 163–164.

46. Tourtouri EN, Delogu F, Sikos L, Crocker MW. Rational over-specification in visually-situated comprehension and production. *J. Cult. Cogn. Sci*. 2019. doi:10.1007/s41809-019-00032-6.

47. Tyrsin AN, Sokolova IS. Entropy-probabilistic modeling of Gaussian stochastic systems. *Matem. Mod.* 2012;24(1):88-102.

48. van Berkum JJA, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn*. 2005;(31):443-467.

49. van Berkum JJA, Zwitserlood P, Hagoort P, Brown CM. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cogn. Brain Res*. 2003;(17):701-718.

50. Venhuizen NJ, Crocker MW, Brouwer H. Semantic Entropy in Language Comprehension. *Entropy.* 2019;(21(12)):1159.

51. Venhuizen NJ, Crocker MW, Brouwer H. Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Process*. 2019;(56):229-255. doi:10.1080/0163853X.2018.1448677.

52. Wu S, Bachrach A, Cardenas C, Schuler W. Complexity metrics in an incremental right-corner parser. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Stroudsburg, PA, USA; 2010. Pp. 1189–1198.

Информация об авторах

Е. В. Шелестюк — доктор филологических наук, доцент, профессор кафедры теоретического и прикладного языкознания.

Е. А. Щетинкина — аспирант кафедры теоретического и прикладного языкознания.

Information about the authors

E. V. Shelestyuk — Doctor of Philological Sciences, Associate Professor, Professor of the Department of Theoretical and Applied Linguistics.

E. A. Shchetinkina — postgraduate student of the Department of Theoretical and Applied Linguistics.

Статья поступила в редакцию 06.08.2022; одобрена после рецензирования 30.08.2022; принята к публикации 26.12.2022.

The article was submitted 06.08.2022; approved after reviewing 30.08.2022; accepted for publication 26.12.2022.

Вклад авторов: оба автора сделали эквивалентный вклад в подготовку публикации.
Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article.
The authors declare no conflicts of interests.