8. Леонтьев, А.А. Формы существования значения / А.А. Леонтьев // Психолингвистические проблемы семантики. – М., 1983.

9 Попова З.Д., Стернин И.А. Очерки по когнитивной лингвистике.– Воронеж, 2001. – 191 с.

10. Серл, Дж. Открывая сознание заново / Дж. Серл. – М.: Едиториал УРСС, 2002.

ИНФОРМАЦИОННАЯ ЭНТРОПИЯ И ЕСТЕСТВЕННЫЙ ЯЗЫК

А.М. Аматов
(Белгород)

Энтропия – одно из ключевых понятий термодинамики, статистической механики и теории информации, описывающее состояние той или иной системы. Естественный язык, разумеется, не является ни механической, ни термодинамической системой, однако он, безусловно, представляет собой древнейшую из известных нам информационных систем, поэтому распространение на него понятия «информационная энтропия» вполне правомерно.

Вообще-то понятия информации и энтропии тесно и органично связаны друг с другом, однако представления об информационной энтропии появились лишь тогда, когда развитие термодинамики и статистической механики сделали эту связь очевидной. У истоков этого понятия стоит логик и математик К. Шеннон, в честь которого информационную энтропию часто называют «энтропией Шеннона». И прежде чем перейти к понятию «энтропия языка», попробуем разобраться в том, что представляет собою информационная энтропия.

Пользуясь интуитивно понятными выражениями, можно сказать, что информационная энтропия – это степень неопределённости сигнала или, применительно к речи, высказывания (англ. *uncertainty* – термин К. Шеннона). В качестве примера возьмём ящик с одинаковыми по размеру и массе шариками, на которых проставлены разные номера. Аналогичный (и весьма расхожий) пример с шариками разного цвета при ближайшем рассмотрении выглядит не совсем удачным, т.к. при достаточно большом их количестве непросто бывает это образно себе представить – в самом деле, вряд ли кто-то может вообразить себе 1000 разных цветов и оттенков. Итак, будем считать, что в ящике лежат 1000 шариков с номерами от 1 до 1000, а некто случайным образом извлекает их из ящика один за другим. При первой попытке неопределённость номера извлечённого шарика максимальна, т.е. вероятность извлечения любого из шариков одинакова и равна 1/1000. Допустим, при первой попытке был вынут шарик под номером 345. Это значит, что он выпал из системы, и в следующей попытке

участвовать не будет. Соответственно, при втором извлечении вероятность случайного выбора любого другого шарика несколько возрастёт, составив 1/999, и будет увеличиваться при последующих попытках: 1/998, 1/997 и т.д., а энтропия системы будет снижаться, пока не останется последний шарик (скажем, с номером 102) и вероятность его извлечения будет равна 1. Если же после каждого извлечения шарик возвращать обратно в ящик, то энтропия будет сохранять максимальное значение для данной системы, т.к. все варианты из тысячи возможных будут равновероятными. Наконец, если предположить, что в ящике всё те же 1000 шариков, однако 100 из них имеют номер 100, а далее – по нарастающей от 101 до 1000, то энтропия системы не будет максимальной, поскольку при первой попытке результаты не будут равновероятными: вероятность извлечь шарик с номером 100 будет существенно выше (1/10), чем у любого другого шарика (1/1000).

Отметим основные характеристики информационной энтропии системы:

- Если все возможные результаты в заданной системе имеют одинаковую вероятность (как извлечение шариков с номерами 1 – 1000 из описанного выше примера), то энтропия системы максимальна.
- Если вероятность какого-либо результата равна 1 (результат точно определён), то энтропия системы равна 0.
- Изменение вероятности события на определённую величину изменяет количество энтропии также на определённую величину.

К. Шеннон [6] формализовал эти положения и выразил информационную энтропию системы через дискретную переменную X, у которой возможен ряд состояний $x_1 \ldots x_n$ в следующей математической формуле:

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_2 \left(\frac{1}{p(x_i)} \right) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (1),$$

где $p(x_i)$ – вероятность i-того события в системе X.

Позднее К. Шеннон пытался применить понятие энтропии и к естественному (английскому) языку [5], однако дальше него пошёл советский математик А.Н. Колмогоров, применивший понятие энтропии Шеннона (вместе с соответствующей формулой её расчёта) к значительному количеству художественных текстов [1; 2; 3 и др.]. Работу в этом направлении продолжил В.А. Успенский, опубликовавший «Семиотические послания» А.Н. Колмогорова (сопроводив их внушительными комментариями и вступительной статьёй) в журнале «Новое литературное обозрение» [5]. Именно Колмогоров ввёл в обиход термин «энтропия языка» и даже рассчитал её. Казалось бы, чего же больше? Явление изучено и даже математически описано. Однако попробуем не спеша разобраться как с работами А.Н. Колмогорова и В.А. Успенского, так и с тем, как энтропия и её рас-

чёт могут найти своё применение в современной лингвистике.

Прежде всего, в работах и А.Н. Колмогорова, и В.А. Успенского смущает неточность некоторых формулировок, вплоть до откровенной некорректности. Разумеется, это не упрёк двум математикам: в конце концов, им простительно и перепутать такие понятия, как язык, речь и текст. Однако то, что может позволить себе в лингвистике математик, не должен допускать языковед. Так, А.Н. Колмогоров [3, 4] пишет: *Вполне естественным является чисто комбинаторный подход к понятию «энтропии речи», если иметь в виду оценку «гибкости» речи – показателя разветвлённости возможностей продолжения речи при данном словаре и данных правилах построения фраз. Для двоичного логарифма числа **N** русских печатных текстов, составленных из слов, включённых в Словарь русского языка С.И. Ожегова и подчинённых лишь требованиям «грамматической правильности», длины **n**, выраженной в «числе знаков» (включая «пробелы»), М. Раттер и Н. Светлова получили оценку **h** = 1,9 ± 0,1. Это значительно больше, чем оценки сверху для «энтропии литературных текстов», получаемые при помощи различных методов «угадывания продолжений». Такое расхождение вполне естественно, так как литературные тексты подчинены не только требованию «грамматической правильности».* Во-первых, непонятно, об энтропии чего сказано в этом отрывке. Вначале вроде бы об энтропии речи, но тут же о *«данном словаре и данных правилах построения фраз»*, т.е. уже как бы и о языке. А в следующем предложении автор почему-то начинает писать о русских печатных текстах с заданной длиной (т.е. количеством *«знаков, включая пробелы»*). Сам собою возникает интересный вопрос: а если какой-либо народ не имеет письменности, то какие он может создавать тексты *«длины **n**, выраженной в «числе знаков»»*? Язык у такого народа есть, но нет письменности. Как же мы будем считать энтропию такого бесписьменного языка? По Колмогорову – никак.

Также и у В.А. Успенского [4, 163] читаем: *Пусть энтропия языка равна **H**. Тогда существует примерно **2Hk** текстов длины **k**, принадлежащих данному языку. Отсюда следует, что чем более узкий корпус текстов мы соотносим с представлением о языке, тем меньше будет энтропия языка; так, если взять энтропию языка русской художественной литературы или энтропию языка русского ямба, то каждая из них будет меньше энтропии русского языка в целом.* Но простите, что значит *«тексты, принадлежащие данному языку»*? Текст – это речевое произведение, которое строится в соответствии с правилами языка, но не *«принадлежит ему»*. Непонятно также, почему энтропия языка зависит от того, какой *«корпус текстов»* мы с этим языком соотносим. Если мы соотнесём с русским языком один лишь текст сказки про колобка, а В.А. Успенский – полное собрание сочинений всех русских классиков, то результаты, со-

гласно Успенскому, должны получиться разными: у нас значение энтропии русского языка будет меньше, чем у него. Но ведь это мы тексты рассматривали разные, а язык как система – один. Более того, в абсолютном значении получается, что энтропия языка тем выше, чем больше на нём написано текстов – весьма спорный вывод. И вообще, что такое *«язык русского языка»*?

Так о какой же энтропии идёт речь в работах Колмогорова, Успенского и других исследователей, применяющих формулу Шеннона к печатным текстам, состоящих из *n* символов? Разумеется, не об энтропии языка, а об энтропии текста, построенного в соответствии с заданными правилами орфографии (которые, строго говоря, непосредственно к языку не относятся). Статистический анализ числа N таких текстов может дать представление об энтропии орфографической системы – и всё. К энтропии же языка как *знаковой* системы такой анализ не приближает нас ни на шаг.

Что же следует учитывать при расчёте энтропии языка? Разумеется, языковой, а не печатный знак. Во-первых, как уже говорилось выше, язык вполне может существовать и без письменности (когда-то все языки обходились без неё), не переставая при этом быть системой со своим уровнем энтропии. Во-вторых, буквы вторичны даже не по отношению к собственно языковым знакам, а к дознаковым единицам языка – фонемам. Соответственно, в восприятии и распознании речи основная нагрузка ложится на те единицы, которые манифестируют именно языковые знаки: морфемы, слова, фразы предложения. В качестве примера можно взять такой текст небольшой газетной заметки:

По рзелульаттам илссеовадний одонго анлигйсокго унвиертисета, не иеемт занчнеия, в кокам пряокде рсапожолены бкувы в солве. Галвоне, чотбы преавя и пслоендяя бквуы блыи на мсете. Осатьлыне бкувы мгоут селдовтаь в плооим бсепордяке, всё рвано, ткест чтаитсея без побрелм. Пичриоий эгото ялвятеся то, что мы не чиатем кдаужю бкуву по отдльснотси, а всё солво цликеом. **Галвный радектор.**

Как показывает данный отрывок, текст действительно «чтаитсея без побрелм», несмотря на «плоный бсепордяок» в буквах. Таким образом, если мы хотим выяснить уровень энтропии языка, то и рассматривать надо именно языковой знак – единство означающего и означаемого.

Вообще, говоря о языке, следует постоянно помнить о том, что это незамкнутая система. Можно сказать, что язык получает «подпитку» энергией извне, поскольку взаимодействует с другими системами (языками, обществом) и здесь вопрос уже выходит за рамки языкознания. Нам в этой связи стоит отметить, что энтропия в системе языка вовсе не обязательно должна нарастать, как это бывает, например, в замкнутых термодинамических системах. Соответственно, энтропия применительно к языку показывает уровень беспорядка при порождении и/или интерпретации

высказывания с учётом фонетики, словаря и грамматических правил.

Далее, в вопросе об энтропии в языковой системе возникает необходимость определения понятий порядка и беспорядка. Безусловно, беспорядок не следует понимать в повседневном смысле слова. В повседневности этот термин имеет весьма размытое значение, поскольку нет и чёткого определения того, что такое порядок. Ну а без чёткого определения порядка невозможно определить и беспорядок.

Применительно к современному языкознанию ситуация напоминает скорее повседневность, нежели научную точность. Действительно, как определить, что в системе языка следует считать порядком, а что беспорядком? Говоря о микросостояниях системы, мы попросту не имеем никаких сколько-нибудь строгих критериев упорядоченности.

Проблему порядка и беспорядка в системе языка в принципе можно решить, если учесть, что беспорядок возрастает с уменьшением вероятности конкретного события. Скажем, беспорядок при бросании кости (6 событий, вероятность каждого – 1/6) выше, чем при бросании монеты (2 события, вероятность каждого – 1/2). С языком всё обстоит значительно сложнее, и не только потому, что количество рассматриваемых событий существенно выше, но и потому, что взаимодействие языка с другими системами подразумевает непрерывное вмешательство извне, как если бы у бросаемой кости кто-то поочерёдно делал ту или иную сторону тяжелее, тем самым увеличивая вероятность конкретного события, впрочем, никогда не доводя её до 1.

Итак, в уравнении (1) необходимо, прежде всего, определить значение термина x_i. По сути, это должен быть показатель уровня неопределённости языкового знака. Применительно к живому языку, мы можем выразить этот показатель через отношение суммы планов содержания к сумме планов выражения, зафиксированных в языке на тот или иной момент времени, или

$$U = \frac{\sum C}{\sum F}.$$

Для какой-либо подсистемы языка, состоящей из ***n*** **элементов,** имеющих m значений,

$$U = \frac{\sum_{i=1}^{m} C_i}{\sum_{j=1}^{n} F_j},$$

где U – показатель неопределённости языкового знака (от *uncertainty*), C – план содержания (от *content*), а F – план выражения (от *form*). В диахронии же уместно будет рассмотреть динамику роста показателя U, т.е. $\frac{dU}{dt}$.

Теперь, если подставить U в формулу (1), мы увидим, что при $U > 1$ энтропия языка будет больше 0 ($H(L) > 0$), при $U = 1$ энтропия будет нулевой ($H(L) > 0$), а при $U < 1$ энтропия будет отрицательной ($H(L) < 0$).

Язык, в котором одному плану содержания соответствует один и только один план выражения (энтропия равна 0), следует считать идеально упорядоченным языком. Если отвлечься от естественного языка, то можно заметить, что в искусственных знаковых системах энтропию часто стремятся свести к нулю. Скажем, система дорожных знаков – это тоже своеобразный язык, в котором каждому знаку соответствует строго одно чтение. Если бы знак можно было трактовать по-разному, это было бы чревато неприятными ситуациями на дороге (которые и без того не редкость). Язык же, в котором одному плану содержания соответствует более одного плана выражения (энтропия отрицательна) будем считать избыточно упорядоченным. Здесь с примерами несколько труднее, но можно вспомнить денежную систему, в которой (обычно так бывает ограниченное время) имеют хождение разные денежные знаки с одним и тем же номиналом: скажем, старые и новые стодолларовые купюры, обычные и «юбилейные» монеты и т.п. Тут как раз и получается, что двум (а возможно и более) планам выражения (вид монет или купюр) соответствует один план содержания (количество товаров и услуг, которые можно на эту купюру приобрести). Разумеется, для конкретного естественного языка сложно вычислить точный показатель энтропии, но сложно – не значит невозможно. Пока же сделаем интуитивное предположение, что показатель энтропии любого естественного языка выше 1, и, скорее всего, такое предположение будет правильным. Для иллюстрации правильности (или, по крайней мере, непротиворечивости) нашей гипотезы из всех подсистем языка удобнее всего рассматривать лексику. Так, в идеально упорядоченном языке одному слову соответствует строго одно лексическое значение, но если мы откроем любой словарь, то обнаружим, что дело обстоит совершенно иначе. Полисемия распространена повсеместно, и отношение количества слов к количеству выражаемых ими значений – один из аспектов общей энтропии языка. Действительно, энтропия всей системы не может снижаться, если растёт энтропия её подсистем.

С грамматикой всё обстоит несколько сложнее, поскольку здесь есть определённые трудности с установлением количества планов содержания. Однако и здесь нередки случаи морфологического и синтаксического гомоморфизма, который в чём-то сродни лексической омонимии. Рассмотри несколько английских предложений:

(2) a. *Bob and Sally are visiting relatives.*
b. *Hunting tigers can be dangerous.*
c. *Sam forgot how good beer tastes.*
d. *We discussed the movie with Bruce Willis.*

В образцах (2 a – d) мы имеем 4 плана выражения на 8 планов содержания, причём ни в одном из приведённых примеров нет случаев лексической омонимии – корни слов, входящих в состав всех этих предложений, сохраняют своё значение. Дело здесь в различии глубинных синтаксических структур.

Конечно, можно сказать, что «речь всё расставит по местам», то есть многозначность в языке, как правило, оборачивается однозначностью речи, однако и здесь есть свои нюансы. Ну, например, многие фигуры речи, такие, как каламбур или зевгма, как раз построены на различии планов содержания при тождестве планов выражения. К слову, при идеально упорядоченном языке такие фигуры были бы невозможны. И потом, не следует забывать, что речь – она только тогда речь, когда привязана к конкретной ситуации, а речевая ситуация во многом случайна, и то, как она будет разворачиваться, предсказать бывает очень сложно. Мало разве в жизни бывает случаев, когда одни люди неправильно толкуют речь других, даже в совершенстве владея языком. Здесь можно сказать лишь то, что чем выше энтропия в системе языка, тем выше степень беспорядка в речи, построенной в соответствии с правилами этого языка.

Безусловно, уровень энтропии различается по языкам. В частности, для современного английского языка этот показатель должен быть несколько выше, чем для русского. Взять, к примеру, предложения (2 a – c). Здесь при переводе каждого из значений на русский язык мы получим отдельное высказывание: *Боб и Салли – гостящие родственники* и *Боб и Салли гостят у родственников*; *Охота на тигров может быть опасной* и *Охотящиеся тигры могут быть опасны*; *Сэм забыл, какой хороший вкус у пива* и *Сэм забыл, какой вкус у хорошего пива*. Что касается предложения (2 d), то оно сохраняет двойственность и при переводе на русский язык: *Мы обсудили фильм с Брюсом Уиллисом*. Без соответствующих пояснений может быть непонятна роль Брюса Уиллиса – либо он сыграл в фильме, либо участвовал в обсуждении (а может быть, и то, и другое вместе).

В этой связи представляется возможным один способ преодоления влияния энтропии языка на алгоритмы порождения и интерпретации речи. Суть его заключается в следующем: формализация лексикона путём задания каждому его элементу (т.е. лексеме) определённого набора грамматических категорий. Таким образом, создаётся что-то вроде словаря, в котором вместо лексико-семантических вариантов того или иного слова представлена номенклатура его синтаксических категорий.

Наибольшую сложность и важность представляет собой синтаксическая категоризация глаголов, поскольку именно глагол выполняет основную синтаксическую функцию, входя в ядро предиката. Для глаголов такие квантитативные категории определяют количество и состав актантов, взаимодействуя с которыми он образует синтаксические структуры. При присвоении глаголу

той или иной синтаксической категории учитываются трансформационные возможности предиката, но не его внутренняя семантика.

При таком подходе получается, что один и тот же глагол обладает некоторым набором синтаксических категорий, образующих парадигму и реализуемых в конкретном типе синтаксической конструкции. Лабильность же с точки зрения трансформационного потенциала следует рассматривать как способность глагола вступать в синтаксические связи, как допускающие пассивизацию, так и не допускающие её. Наконец, наличие у одного глагола ряда квантитативных категорий не может представлять особой проблемы (например, стоит ли рассматривать эти единицы как один глагол или как разные). Ведь, в конце концов, наличие у одной и той же единицы ряда лексико-семантических вариантов ни у кого нареканий не вызывает.

Литература

1. Колмогоров, А.Н. Пример изучения метра и его метрических вариантов // Теория стиха. АН СССР: Ин-т русской литературы (Пушкинский дом). – Л.: Наука, 1968. – С. 145-167.

2. Колмогоров, А.Н. Теория информации и теория алгоритмов – М.: Наука, 1987.

3. Колмогоров, А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – № 1, 1965. – С 3-11.

4. Успенский, В.А. Предварение для читателей «Нового литературного обозрения» к семиотическим посланиям Андрея Николаевича Колмогорова // Новое литературное обозрение. – № 24, 1997.

5. Шеннон, К. Работы по теории информации и кибернетике. Пер. с англ. Предисл. А. Н. Колмогорова. – М.: Изд-во иностранной литературы, 1963.

6. Shannon, C.E. A Mathematical Theory of Communication // Bell System Technical Journal. – Vol. 27, July and October, 1948. – Pp 379-423 and 623 – 656

ФЕНОГРАММАТИЧЕСКИЕ ФОРМЫ И ТЕКТОГРАММАТИЧЕСКИЕ СТРУКТУРЫ (на примере английских сложных предикатов с прилагательным и инфинитивом)

О.А. Аматова
(Белгород)

Термины «фенограмматическая форма» и «тектограмматическая структура» представляют собой дальнейшее развитие понятий порождающей грамматики Н. Хомского «поверхностная структура» и «глубинная структура» в рамках трансформационных категориальных грамматик GPSG и HPSG (см., напр., [1]). Фенограмматическая форма в сущности