Эксперименты по непредсказуемости слова в контексте и поток информации в естественном языке*

Д.Ю.Манин

6 ноября 2006 г.

Аннотация

Данные крупномасштабного эксперимента показывают, что логарифм вероятности угадать слово в отрывке текста (непредсказуемость) линейно зависит от длины этого слова. Этот результат справедлив как для поэзии, так и для прозы, несмотря на то, что в прозе испытуемые не могут определить длину пропущенного слова. Выдвинута гипотеза о том, что этот эффект отражает тенденцию естественного языка к поддержанию равномерности потока информации.

1 Введение

Настоящая работа посвящена одному частному результату экспериментального исследования предсказуемости слов в контексте. Основная цель эксперимента — исследование некоторых аспектов восприятия поэзии, но излагаемый здесь результат по нашему мнению, представляет общелингвистический интерес.

Первое исследование предсказуемости естественного текста было предпринято основателем теории информации Клодом Шенноном [1]. (Заметим, что в своей основополагающей работе [2] Шеннон затронул и вопрос о соотношении между литературными качествани текста и его теоретико-информационной избыточностью, сравнив высокоизбыточный т.н. "базовый английский"с "Поминками по Финнегану"Джойса, где "расширение словаря приводит, как утверждается, к повышенной концентрированности семантического содержания".) Шеннон предъявлял испытуемому (своей жене) случайно выбранные отрывки из биографии президента Джефферсона и предлагал угадывать следующую букву до тех пор, пока не получал правильный ответ. Количество попыток до правильного ответа использовалось для вычисления верхней и нижней оценок энтропии английского языка,

*Эта статья представляет собой несколько расширенный русский текст работы Manin, D.Yu. 2006. *Experiments on predictability of word in context and information rate in natural language.* J. Information Processes (electronic publication, http://www.jip.ru), 6 (3), 229-236

которая оказалась в пределах между 0.6 и 1.3 бита на знак (б/з), что значительно меньше энтропии случайной последовательности равновероятных букв латинского алфавита. Шеннон показал также, что условная энтропия убывает по мере увеличения длины отрезка предшествующего текста, предъявляемого испытуемому, по крайней мере пока она не превышает 100 знаков.

Несколько авторов повторяли и усовершенствовали эксперимент Шеннона. Бертон и Ликлидер [3] использовали 10 разных текстов сходного стиля и варьировали длину фрагментов от 1 до 1000 знаков. Они заключили, что, в отличие от данных Шеннона, увеличение длины предъявляемого отрывка свыше 32 знаков не уменьшает энтропию.

Фонодь [4] сравнил предсказуемость следующей буквы для трех типов текста: стихов, газетной статьи и "разговора двух девушек". Его методика, по-видимому, предусматривала только одну попытку угадывания на букву, поэтому оценки энтропии получить из этих результатов нельзя (см. ниже). Автор сообщает результаты в форме частоты правильных ответов, причем стихи оказываются значительно менее предсказуемы, чем тексты двух других типов.

Колмогоров в своей работе [5], заложившей основы алгоритмической теории сложности ("колмогоровская сложность") приводит результаты своих экспериментов, давших оценку в 0.9–1.4 б/з. Подробности (чрезвычайно остроумной) экспериментальной методики в работе не приводятся, но она описана в известной монографии А.М. и И.М. Ягломов "Статистика и теория информации".

Кавер и Кинг [6] модифицировали методику Шеннона другим образом, предлагая испытуемым делать денежные ставки на следующую букву текста. Они продемонстрировали, что при оптимальной стратегии игроки будут распределять наличный капитал между возможными исходами испытания пропорционально оценке их вероятности. Поэтому в предположении об оптимальной игре (вообще говоря, небесспорном) можно оценить эти вероятности, исходя из делаемых ставок. Эта методика дала результат в 1.3 б/з для английского языка. В этой работе также содержится обширная библиография.

Моради с соавторами [7] сначала использовали два разных текста (учебник по цифровой обработке сигналов и роман Джудит Кранц) для подтверждения результата Бертона и Ликлидера о критической длине фрагмента (32 знака), а затем добавили еще два ("101 далматин"и федеральный авиационный справочник) для изучения зависимости энтропии от типа текста и ее вариации по испытуемым (с несколько неопределенными результатами).

Множество работ было посвящено оценке энтропии естественного языка с помощью статистиеских методов, без использования подопытных субъектов. Один из первых таких опытов описан в [8], где 39 английских переводов 9 классических греческих текстов использовались для изучения зависимости энтропии от темы, стиля и периода написания. При этом, однако, использовалась очень грубая оценка, основанная на частоте двухбуквенных сочетаний. Некоторых из более поздних результатов в этом направлении можно найти в [9], [10] и цитированной там литературе. По самой своей природе подобные

методы не используют семантику (и даже синтаксис) текста, но за счет вычислительных мощностей современных компьютеров они достигают результатов (вероятности угадывания следующей буквы), до некоторой степени сравнимых с человеческими.

Постановка нашего эксперимента отличается от предыдущих работ в двух принципиальных отношениях. Во-первых, наши испытуемые угадывают целые слова, а не отдельные буквы. Во-вторых, для угадывания предлагаются слова из любого места отрывка, а не только последнее. Кроме угадывания пропущенного слова, испытуемым предлагаются два других типа задания, где оценивается подлинность предъявляемых слов. Причина этих отличий в том, что в то время, как предшествующие исследования были мотивированы преимущественно разработкой методов эффективного сжатия текстов, нас интересуют литературные тексты именно с точки зрения их литературности, а не как подлежащие сжатию цепочки символов[1]. Нашей целью при разработке эксперимента было получить исходные данные для проверки некоторых гипотез в области теоретической поэтики, которые без этого остаются неизбежно спекулятивными. Угадывать следующее слово — не очень адекватная задача для литературного текста, поскольку даже обыкновенное повествовательное предложение (например, это) является, по существу, не линейной цепочкой символов или слов, но сложной структурой, пронизанной связями между словами, как прямыми, так и обратными. Стихотворение же — в еще большей степени структура из существенно взаимосвязанных элементов, которая не читается последовательно, и уж конечно, не пишется последовательно. Кроме того, практика показывает, что угадывая букву за буквой, люди на самом деле всегда основывают свой выбор на предположении о следующем слове. По всем этим причинам для наших целей более уместно было работать со словами, а не буквами.

Однако поскольку представленные здесь результаты, как уже упоминалось, не относятся собственно к поэтике, мы не задерживаемся более на этом и отсылаем заинтересованного читателя к работе [11].

2 Постановка эксперимента

Во Введении к специальному выпуску журнала по вычислительной лингвистике с использованием больших корпусов текстов Черч и Мерсер [12] отмечают: "90-е годы ознаменовались возобновлением интереса к эмпирическим и статистическим методам анализа языка в стиле 50-х". Они относят этот ренессанс эмпиризма преимущественно на счет возросших вычислительных мощностей и появления огромных массивов оцифрованных данных. Конечно, эти факторы способствуют применению статистического анализа текстов как символьных последовательностей. Однако широкая доступность компьютерных сетей и

[1] Следует отметить, что эффективное сжатие представляет интерес не только само по себе, но и для криптографических приложений, как отмечено в [10]. Кроме того, языковые модели, разработанные для целей сжатия, успешно используются также и в таких приложениях, как распознавание речи и печатного текста, позволяя исправлять ошибки и разрешать неопределенности в трудных местах.

интерактивные технологии позволяют также ставить эксперименты с участием людей в немыслимых раньше масштабах.

Описываемый в настоящей работе эксперимент поставлен в форме сетевой литературной игры. Добровольным участникам игры доступна также вся информация о ее исследовательской стороне, включая основные текущие результаты. Они также могут принимать участие в обсуждениях.

Игрокам предъявляются фрагменты текстов, в которых одно из слов заменено пробелами или другим словом. "Словом"считается любая последовательность из 5 или более русских букв между не-буквами. Слова выбираются для предъявления случайно и равновероятно. Предусмотрено три типа заданий:

тип 1: угадать опущенное слово;

тип 2: определить, является ли авторским выделенное слово;

тип 3: определить, которое из двух предъявленных слов является авторским.

Неправильные ответы на задания типа 1 используются в качестве замен в заданиях двух других типов.

Фрагменты выбираются случайно из корпуса в 3439 отрывков главным образом стихотворных текстов в широком диапазоне стилей и времен: от авангардных до популярных, от классических до любительских и от XVIII века до современных. Для сравнения включены также три прозаических текста ("Анна Каренина", "Доктор Живаго"и современное публицистическое эссе).

К настоящему времени (весна 2006 г.) данные набирались почти непрерывно в течение трех лет. Более 8000 человек приняли участие в эксперименте, совершив в общей сложности почти 900 тыс. угадываний, примерно треть из которых относится к типу 1. В традиционном лабораторном эксперименте подобные масштабы недостижимы. Разумеется, методика не свободна от недостатков, в первую очередь благодаря неконтролируемости условий эксперимента. Эти вопросы подробно разбираются в [11]. Но недостатки с лихвой окупаются статистическим весом данных, особенно если учесть, что иным способом достичь его не представляется возможным.

3 Результаты

Непосредственной целью эксперимента было изучение систематических различий между разными категориями текстов с точки зрения того, насколько трудно в них а) восстановить опущенное слово и б) отличить авторское слово от замены. Однако в настоящей работе мы рассматриваем некоторое конкретное свойство текстов, которое оказалось независящим от типа текста и таким образом, по-видимому, характеризует не текст, а язык в целом. Это свойство — зависимость непредсказуемости слова от его длины.

Определим *непредсказуемость* U как отрицательный двоичный логарифм вероятности угадать слово, $U = -\log_2 p_1$, где p_1 — средняя частота правильных ответов на задания типа 1. Для отдельного

слова это определение формально эквивалентно шенноновскому определению энтропии H. Однако если объединить разные слова, энтропия должна вычисляться как средний логарифм вероятности, а не как логарифм средней вероятности:

$$H = -\frac{1}{N}\sum_{i=1}^{N}\log_2 p_1^i \tag{1}$$

$$U = -\log_2 \frac{1}{N}\sum_{i=1}^{N} p_1^i \tag{2}$$

Действительно, логарифм вероятности угадать слово дает количество информации в битах, теоретически достаточной для того, чтобы зафиксировать выбор этого слова. Таким образом, усреднению подлежит именно эта величина. Практически для оценки вероятности используют наблюденную частоту. Однако в реальном эксперименте всегда есть слова, не угаданные ни разу, для которых эта оценка вероятности дает $p_1 = 0$ и неопределенное значение логарифма (именно поэтому методика Шеннона предусматривает угадывание до правильного ответа). Формально, если в последовательности есть элемент, вероятность угадать который равна нулю (или очень мала), то количество информации во всей последовательности может определяться одним этим элементом.

С другой стороны, значение определенной выше непредсказуемости чувствительно не к точной вероятности угадать "трудные"слова, а только к тому, какова их доля в тексте. Если энтропия характеризует количество попыток, необходимых для угадывания случайно выбранного слова, непредсказуемость характеризует долю слов, которые угадываются с первой попытки. Разумеется, эти две величины совпадают, если энтропия всех слов в тексте одинакова.

Проблему, возникающую в связи с неугадываемыми словами, можно обойти, приписав им некоторую произвольно фиксированную энтропию. Мы сравнили непредсказуемость с энтропией, вычисленной в таком приближении с двумя значениями параметра: 10 бит (что примерно соответствует угадыванию вслепую с помощью частотного словаря) и 3 бит (оптимистическая нижнаяя граница). В обоих случаях вычисленное H, не будучи численно равным U, оказалось связанным с ней монотонной, приближенно линейной зависимостью. Это, вероятно, означает, что доля трудноугадываемых слов возрастает и убывает вместе с непредсказуемостью остальных слов. Поэтому мы предпочитаем работать с непредсказуемостью, не вводя произвольных гипотез, вместо того, чтобы вычислять значение энтропии, ценность которого была бы сомнительной.

Непредсказуемость как функция длины слова, вычисленная усреднением по всем словам одной длины и по всем текстам, показана на фиг. 1 и фиг. 2 (где длина слова измеряется в знаках и в слогах соответственно). Доверительные интервалы на графиках соответствуют стандартному отклонению биномиального распределения (поскольку данные являются результатом ряда независимых испытаний с двумя

возможными исходами в каждом: ответ может быть правильным или неправильным).

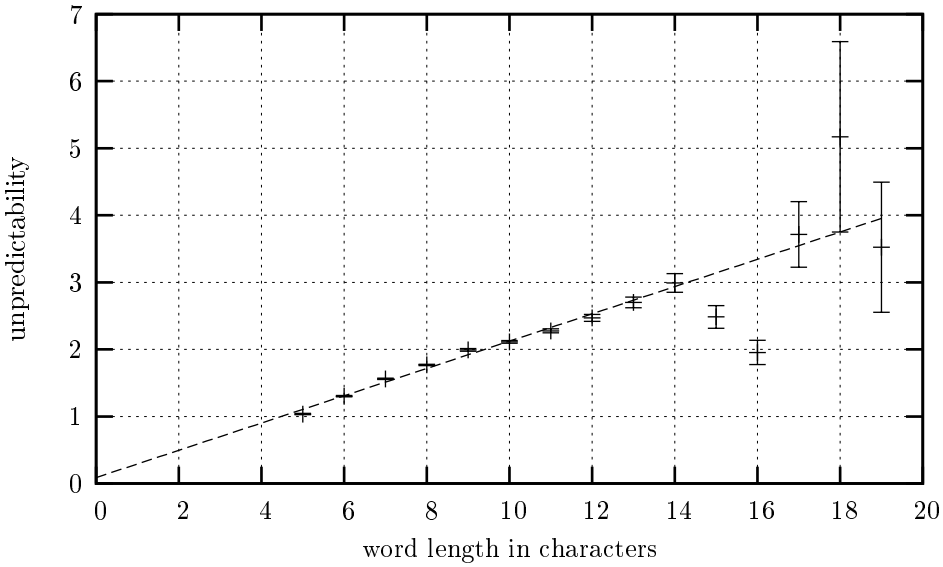

Рис. 1: Непредсказуемость как функция длины слова в знаках, по всем текстам.

В диапазоне от 5 до 14 знаков и от 1 до 5 слогов наблюдается убедительная линейная зависимость. Более длинные слова редки, поэтому данные для них значительно менее надежны статистически. Мы обсуждаем линейную зависимость только в том диапазоне, где она может считаться надежно установленной.

4 Обсуждение

По указанным выше причинам, сравнить наши результаты с предыдущими работами непросто. Все же, можно указать два аспекта, по которым сравнение возможно. Во-первых, можно грубо оценить влияние угадывания слова в контексте по сравнению с угадыванием следующего слова в последовательности. Вспомним, что Шеннон [1] оценил энтропию английского языка в расчете на слово в приближении нулевого порядка с помощью закона Ципфа, получив значение 11.82 бит на слово (б/сл). Браун с соавторами [9] использовали статистическую марковскую модель языка 2-го порядка (вероятности троек слов) и получили оценку в 1.72 б/з, что дает 7.74 б/сл при средней длине английского слова в 4.5 знаков. Это значит, что знание вероятностей троек слов повышает вероятность угадать следующее слово на $11.82 - 7.74 = 4.08$ б/сл. Но слово в контексте участвует сразу в трех тройках: как первое слово в одной, среднее в другой и последнее в

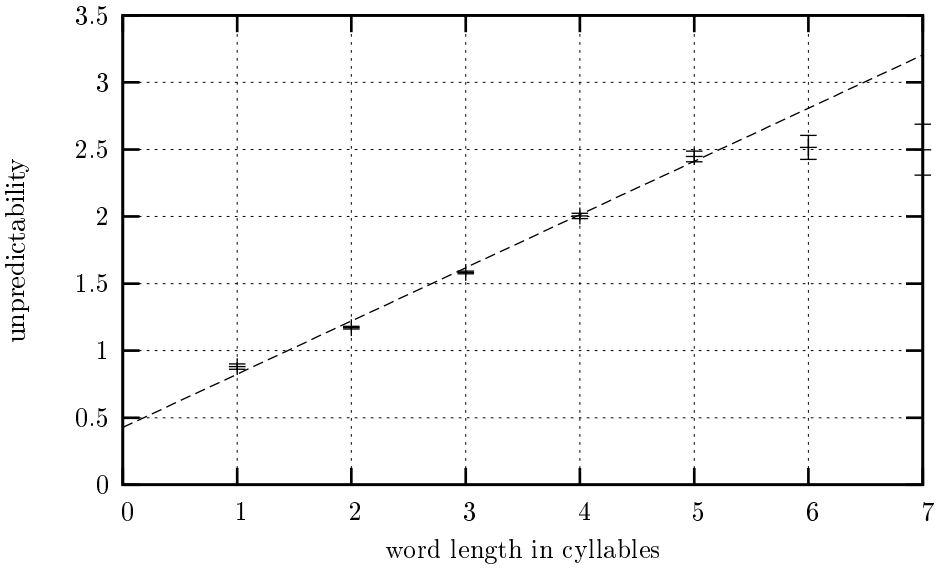

Рис. 2: Непредсказуемость как функция длины слова в слогах, по всем текстам.

третьей. При угадывании следующего слова доступна только первая тройка, а при угадывании слова в контексте — все три (этот гипотетический эксперимент реально не ставился, насколько нам известно). Разумеется, тройки нельзя считать статистически независимыми, поэтому в качестве грубого приближения можно предположить, что последняя тройка содержит несколько меньше информации, чем первая, а средняя — очень мало (поскольку все ее слова участвуют и в двух других тройках). Другими словами, можно ожидать, что такая модель дала бы оценку энтропии примерно на 4 б/сл ниже для слова в контексте, чем для следующего слова в последовательности, что весьма существенно (вероятность угадать повышается на порядок).

Сравнение в другой плоскости возможно с графиком фиг. 13 из работы [13], где отложено значение энтропии для n-й буквы слова в зависимости от позиции n. Эта энтропия вычислялась с помощью алгоритма типа Зива—Лемпеля. Известно, что как статистические модели, так и люди хуже всего угадывают первую букву слова, и график позволяет количественно оценить этот эффект: оказывается, что первая буква имеет энтропию 4 бит, далее энтропия быстро убывает до 5-й буквы, после чего остается на удивление постоянной вплоть до 16-й позиции. (Практически идентичная зависимость получается для текста со случайно переставленными словами, что красноречиво характеризует сильные и слабые стороны современных статистических моделей языка.) По этим данным можно восстановить зависимость энтропии слова от его длины как $h_n^{(w)} = \sum_{i=1}^{n} h_i^{(l)}$, где $h_n^{(w)}$ — энтропия слов длины n, а $h_i^{(l)}$ — энтропия i-й буквы в слове. Эта зависимость, спра-

ведливая для языковой модели [13], резко возрастает от слов длиной в 1 букву до 5-буквенных, после чего следует приблизительно линейный пологий участок с наклоном в 0.6–0.7 б/з. Эта картина сильно отличается от нашей фиг. 1, и хотя мы вычисляем непредсказуемость, а не энтропию, эта разница, по-видимому, остается существенной (ввиду замечания о соотношении энтропии и непредсказуемости выше).

На самом деле, наш результат может на первый взгляд показаться тривиальным. В самом деле, согласно одной из теорем, доказанных Шенноном (теорема 3 в [2]), для символьной последовательности, порождаемой стационарным эргодическим источником, почти все подпоследовательности длины n имеют одну и ту же вероятность, экспоненциальную по n: $P_n = 2^{-Hn}$ для достаточно больших длин (здесь H — энтропия источника). Однако это объяснение неприменимо к нашей задаче по нескольким причинам. Даже если оставить в стороне вопрос об эргодичности естественного языка, с формальной точки зрения, условия теоремы требуют длин n достаточно больших, чтобы все возможные знаковые диграммы (двойки букв) встречались по крайней мере по разу. Нечего и говорить, что это гораздо больше длины одного слова. С доугой стороны, если бы эта теорема и была применима, вероятность отгадать слово должна была бы быть порядка P_n, что много меньше наблюдаемых частот правильных ответов. По существу, наши испытуемые могут угадывать слова только потому, что эти слова связаны с контекстом, осмысленны в нем, в то время как в условиях теоремы Шеннона равновероятные подпоследовательности, наоборот, асимптотически независимы от контекста.

Можно пытаться объяснить линейную зависимость непредсказуемости от длины слова, предположив, что число слов в языке (в словаре или в текстах) заданной длины возрастает с увеличением этой длины, отчего более длинные слова труднее угадывать просто потому, что расширяется набор возможных ответов. Если бы это расширение словаря было экспоненциальным по длине слова, можно было бы предположить, что контекстные ограничения на выбор слова сокращают число возможностей в фиксированное в среднем число раз, так что экспоненциальный характер роста сохраняется и при учете семантики и синтаксиса. Однако данные не согласуются с таким объяснением. Распределение слов по длине, вычислять ли его по реальным текстам или по частотному словарю (мы использовали частотный словарь русского языка, содержащий 32000 слов [14]) не только не экспоненциальное, но даже не монотонное. Число различных слов заданной длины возрастает до длины ок. 8 знаков, после чего убывает. Такое поведение никак не отражается на графиках фиг. 1 и 2, из чего можно заключить, что количество различных слов заданной длины не влияет на угадывание ни в какую сторону.

На самом деле, распределение слов по длине могло бы влиять на угадывание только если бы испытуемые знали длину угадываемого слова. Но это, вообще говоря, не так. В заданиях не дается никаких указаний на длину слова. Поскольку русские стихи по большей части метрические, слоговую длину слова в стихах определить, как правило, нетрудно. Но замечательно, что зависимость непредсказуемости от длины в стихах и в прозе совершенно одинаковая (см. фиг. 3), хотя

в прозе никакой возможности определить длину угадываемого слова (не угадывая само слово) нет. [2]

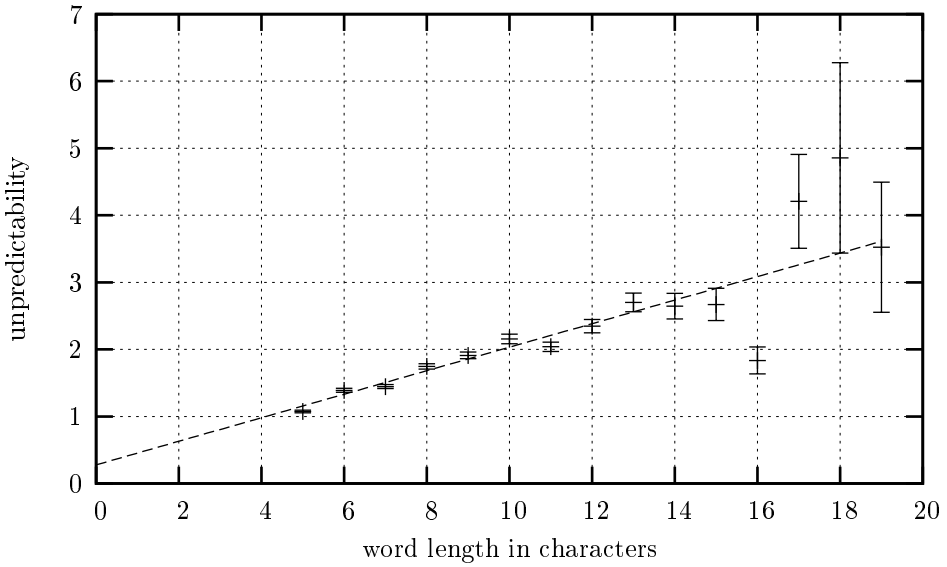

Рис. 3: Непредсказуемость как функция длины слова в знаках, только по прозе.

Таким образом, у нас остается только одно разумное объяснение наблюдаемого эффекта: язык изменяется таким образом, чтобы сглаживать средний поток информации, так что более длинные слова несут пропорционально больше информации. Это было бы естественно, поскольку неравномерность потока информации неэффективна: там, где он меньше оптимального, недоиспользуется пропускная способность канала связи, а там, где превышает оптимальную величину, теряется помехоустойчивость. Иными словами, по мере того, как язык меняется с течением времени, слишком длинные слова и грамматические формы будут укорачиваться, а слишком длинные — удлиняться и усиливаться. Разумеется, поскольку язык меняется под давлением множества разных факторов, оптимум никогда не достигается и существует только как зависящий от времени аттрактор.

Интересно отметить, что подобная гипотеза была мимоходом высказана Черчем и Мерсером в другой связи в цитированной работе [12]. Обсуждая приложения статистических марковских моделей второго порядка, они пишут (стр. 12):

[2]Интересно, что по средней непредсказуемости слов поэзия и проза отличаются удивительно мало. Оказывается, что в поэзии повышению предсказуемости за счет метра и рифмы противодействует ее понижение за счет большей свободы семантики (тропы) и, возможно, синтаксиса. Замечательно, что эти две тенденции практически компенсируют друг друга. Это явление и его значение подробно обсуждаются в [11].

> Вообще, частотные вспомогательные слова вроде *to* и *the*, которые акустически коротки, более предсказуемы, чем более длинные содержательные слова вроде *resolve* и *important*. Это удобно для распознавания речи, поскольку это означает, что языковая модель сильнее ограничивает выбор как раз тогда, когда акустической модели приходится особенно туго. Возникает подозрение, что это не случайность, а естественный результат эволюции речи под давлением человеческой потребности в надежной связи при наличии шума.

В самом деле, было бы удивительно если бы у естественного языка не оказалось свойства, "удобного для распознавания речи". Из наших результатов следует, что оно присуще языку даже в гораздо большей степени, чем можно было бы предположить, исходя из наблюдения Черча и Мерсера. Конечно, это только один из механизмов изменения языка, и действует он только статистически, так что в любом отдельно взятом языковом состояним будут наблюдаться явления с существенно пониженной и существенно повышенной избыточностью. Так, любой носитель русского языка знает, как нелегко отличить "мне надо"от "не надо". В американском английском наблюдались цепные сдвиги гласных, вызванные тем, что изменение произношения одной гласной приводит к неразличимости многих слов (напр., "pin"и "pen"в современном южном диалекте), отчего начинает меняться произношение другой гласной, чтобы восстановить разницу. Такие изменения, вероятно, часто носят характер колебаний. Рассмотрим в качестве примера эволюцию английского отрицания (по книге [15], с. 175–176):

> Исходное староанглийское отрицание выражалось словом *ne*, напр. *ic ne wāt*, 'я не знаю'. Это простое отрицание могло усиливаться гиперболическим добавлением либо слова *wiht* 'что-то, что-либо', либо слова *nāwiht* 'ничто' [...]. С течением времени услительный характер добавления *(nā)wiht* стирался [...] и форма *nāwiht* была переинтерпретирована как часть составного, "разрывного"отрицания *ne ... nāwiht* [...]. Но как только обыкновенное отрицание стало выражаться двухчастной формулой из *ne* и *nāwiht*, возникли условия для выхода на сцену эллипсиса, который устранил то, что воспринималось как избыточность. В результате частица *ne*, которая исходно выражала отрицание, выпала, а слово *not*, потомок исходно гиперболического *nāwiht*, превратилось в единственную отрицательную частицу. (Современный английский испытал дальнейшие изменения благодаря введению "вспомогательного"глагола *do*.)

Это очень напоминает колебания, возникающие при итеративном поиске оптимальной длины данной грамматической формы. Тем более поразительно, как эта тенденция, несмотря на свою статистическую и нестационарную природу, замечательно проявляется в данных.

Добавление к русскому тексту. Уже после того, как эта работа была опубликована, мне стали известны следующие три работы, в которых эффект той же самой природы изучался на других уровнях:

Уровень дискурса. Genzel & Charniak, 2002. *Entropy rate constancy in text.* Proc. 40th Annual Meeting of ACL, 199–206. Рассматривается зависимость энтропии предложения от его позиции в тексте. Показано, что энтропия, вычисленная языковой моделью, и поэтому не учитывающяя семантики, слегка возрастает для не слишком больших позиций. Делается вывод, что с учетом семантики эта энтропия была бы постоянной, потому что содержание предшествующего текста помогало бы предсказывать последующий.

Уровень предложения. Jaeger, 2006. *Speakers optimize information density through syntactic reduction.* To be published. В некоторых английский предложениях подчинительный союз *that* носит необязательный характер. Экспериментально показано, что носители языка чаще употребляют этот союз в тех предложениях, где выше плотность информации, таким образом "разбавляя"их.

Уровень слога. Aylett & Turk, 2004. *The smooth signal redundancy hypothesis [...].* Language and Speech, 47(1), 31–56. Показано, что в речи усиливаются и удлиняются те слоги, которые менее предсказуемы, тем самым поддерживается более равномерный уровень избыточности.

Список литературы

[1] C. E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30:50–64, January 1951.

[2] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[3] N.G.Burton and J.C.R.Licklider. Long-range constraints in the statistical structure of printed english. *American Journal of Psychology*, 68(4):650–653, December 1955.

[4] I. Fónagy. Informationsgehalt von wort und laut in der dichtung. In *Poetics. Poetyka. Поэтика*, pages 591–605. Państwo Wydawnictwo Naukowe, Warszawa, 1961.

[5] Kolmogorov, A. Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1:1–7, 1965.

[6] T.M.Cover and R.C.King. A convergent gambling estimate of the entropy of english. *Information Theory, IEEE Transactions on*, 24(4):413–421, jul 1978.

[7] Hamid Moradi, James A. Roberts, and Jerzy W. Grzymala-Busse. Entropy of english text: Experiments with humans and a machine learning system based on rough sets. *Inf. Sci*, 104(1-2):31–47, 1998.

[8] W.J.Paisley. The effects of authorship, topic structure, and time of composition on letter redundancy in english text. *J. Verbal. Behav.*, (5):28–34, 1966.

[9] Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. An estimate of an upper bound for the entropy of english. *Comput. Linguist.*, 18(1):31–40, 1992.

[10] W. J. Teahan and John G. Cleary. The entropy of english using PPM-based models. In *DCC '96: Proceedings of the Conference on Data Compression*, pages 53–62, Washington, 1996. IEEE Computer Society.

[11] R.G.Leibov and D.Yu.Manin. An attempt at experimental poetics [tentative title]. To be published in Proceedings of Tartu University, 2006.

[12] Kenneth W. Church and Robert L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Comput. Linguist.*, 19(1):1–24, 1993.

[13] T. Schürmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427, 1996.

[14] S.Sharoff. The frequency dictionary for russian. http://www.artint.ru/projects/frqlist/frqlist-en.asp.

[15] H.H.Hock and B.D.Joseph. *Language History, Language Change, and Language Relationship*. Mouton de Gruyter, Berlin, New York, 1996.