ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 1 1960

ПРИКЛАДНОЕ И МАТЕМАТИЧЕСКОЕ ЯЗЫКОЗНАНИЕ

ТЕОРИЯ ИНФОРМАЦИИ И ЛИНГВИСТИКА

Речь как вероятностный процесс

В ряде вопросов лингвистики — и в особенности тогда, когда учет смыслового содержания человеческой речи был бы ненужным бременем, мешающим разумной постановке задачи (например, во многих вопросах передачи речи по каналам связи),— оказывается целесообразным подход к языку как к вероятностному процессу. Язык рассматривается при этом как цепь испытаний, могущих иметь в зависимости от случая разные исходы, причем вероятности отдельных исходов каждого испытания существенно зависят от исхода предшествующих испытаний.

Под «испытаниями» мы будем в основном понимать появления отдельных букв письменного текста; зависимость последовательных испытаний означает здесь, что вероятности появления отдельных букв существенно зависят от предшествующих (и последующих) букв. Так, вероятность появления гласной буквы возрастает, если предшествующая буква была согласной; после буквы *ч* никак не могут следовать буквы *ы*, *я* или *ю*, а скорее всего окажется буква *т* (слово *что*) или одна из гласных *и* и *е* и т. д. Под «испытаниями» можно понимать и определение последовательных слов, фонем, морфем и других лингвистических элементов письменной или устной речи.

Подобный подход к языку ведет свое начало от замечательного русского математика А. А. Маркова старшего. Он проиллюстрировал на примере языка разработанную им теорию вероятностных процессов специального типа, которые теперь называют марковскими процессами[1]. Особое значение такой подход приобрел в последние годы в связи со ставшими уже классическими исследованиями американского математика и инженера К. Шеннона[2], лежащими в основе большого научного направления, получившего название «теория информации». Основной заслугой Шеннона является указание на возможность строгого количественного подхода к вопросам, связанным со всевозможными процессами передачи, хранения и переработки информации, — весьма ценное для лингвистики обстоятельство, поскольку передача информации является основной функцией языка.

Заметим, что подход к речи как к вероятностному процессу требует знания вероятностей появления отдельных лингвистических элементов и их сочетаний. Эти вероятности могут быть определены статистически как частоты; так, например, под вероятностью буквы *а* мы будем понимать отношение числа повторений этой буквы в некотором тексте к общему числу букв текста. Такой подход требует строгой фиксации текста, статистические закономерности которого кладутся в основу теории; текст должен быть достаточно обширным («язык Пушкина», т. е. совокупность всех пушкинских текстов, или «русская газетная и журнальная речь XX в.»— совокупность всех текстов, напечатанных в XX в. в русских газетах и журналах, и т. д.). Мы в дальнейшем будем исходить из «среднего современного литературного языка», понимая под этим, что анализируемый текст состоит из всего напечатанного на данном языке за последние годы. При этом нет необходимости обрабатывать статистически весь этот текст; можно ограничиться небольшими (но разнообразными по характеру) и отбираемыми с использованием разумной статистической методики выборками из него. Любой отрывок напечатанного текста будет по своим статистическим закономерностям близок к «среднему языку». Разумеется, некоторые отклонения от него для специальных текстов возможны: так, например, средняя длина слова в научных текстах окажется больше «среднелитературной», так как научная терминология обычно состоит из длинных слов; очень редкая в русском языке буква *ф* будет чаще обычного встречаться в математической литературе из-за частого повторения слов *функция*, *дифференциал*, *коэффициент* и некоторых других.

---

[1] В своей книге «Исчисление вероятностей» (4-е изд., М., 1924) А. А. Марков подверг статистическому анализу два отрывка из «Евгения Онегина» и «Детских лет Багрова-внука».

[2] C. E. Shannon, W. Weaver, The mathematical theory of communication, Urbana, 1949. Неполный русский перевод см.: К. Шеннон, Статистическая теория передачи электрических сигналов, сб. «Теория передачи электрических сигналов при наличии помех», М., 1953.

От Шеннона идет и составление «моделей фраз», основанное на учете тех или иных статистических закономерностей языка. Так, выписывая ряд русских букв (к числу которых причисляется и «нулевая буква», означающая пробел между словами) «наудачу», или учитывая вероятности появления в тексте отдельных букв, сочетаний из двух букв, из трех букв и из четырех букв, можно прийти к таким, например, «фразам»:

Φ_0: оухеррохьдбщ яыхвщихйжтифвнарфенвщтфрпхгпчькизряс;

Φ_1: т цыяь оерв одног збя енвтша ьуемлолйк;

Φ_2: кая всванный рося ных ковкров;

Φ_3: покак постивленный пот дурноскака наконеппо зно стволовил;

Φ_4: весел враться не сухом и непо и добре.

Здесь каждая следующая фраза ближе к осмысленному русскому тексту, чем предыдущая. Если в такой фразе учитываются вероятности n-буквенных сочетаний, где n больше средней длины слова в данном языке, то эта «фраза» будет состоять из слов данного языка; еще более осмысленной будет она, если n очень велико. Однако при любом n мы будем при разных опытах приходить к разным фразам. Это связано с тем, что статистические связи лишь отсекают противоречащие духу языка конструкции, но не закрепляют речь — они допускают еще множество различных вариантов, выбор которых и позволяет передать словами нужную информацию. Осмысленность речи создается ее неопределенностью, так как благодаря неопределенности речи каждое сообщение является для нас неожиданным, новым.

При всем том совершенно ясно, что степень неопределенности фразы Φ_4 меньше степени неопределенности фразы Φ_0; степень неопределенности осмысленного текста должна быть еще меньше, чем для фразы Φ_4. Какова же степень неопределенности осмысленного текста и какую информацию содержит этот текст? Сама постановка этих бесспорно основных для языкознания вопросов стала возможной лишь после разработки общей «теории информации».

Энтропия и информация вероятностного процесса

Понятие информации тесно связано с неопределенностью исхода какого-либо опыта: определение известного заранее результата не может доставить новой информации. Неопределенность же опыта тесно связана с вероятностями его исходов (которые можно описать как долю или процент отдельных исходов при многократном повторении опыта). Некоторые соображения[3] вынудили Шеннона принять, что степень неопределенности опыта α, имеющего n исходов, вероятности которых равны $p_1, p_2, \ldots, p_n$ (где $p_1 + p_2 \ldots + p_n = 1$), выражается числом

$$H(\alpha) = c(-p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n); \qquad (*)$$

это же число измеряет количество информации, которое мы получим, определив исход α. Величину $H(\alpha)$ Шеннон предложил называть энтропией опыта α; можно показать, что она равна нулю лишь в том случае, если одна из вероятностей $p_1, p_2, \ldots, p_n$ равна 1, а остальные — 0 (т. е. если исход опыта α нам известен заранее); в остальных случаях $H(\alpha) > 0$. Коэффициент пропорциональности c в формуле (*) зависит от основания логарифмов и выбора системы единиц. Чаще всего за единицу измерения неопределенности опыта (и информации) принимается так называемая двоичная единица информации, или бит, равная степени неопределенности опыта с двумя равновероятными исходами, например опыта, состоящего в определении стороны, на которую падает подброшенная монета. (Иногда соответствующую единицу информации называют также «да-нет единицей», поскольку она равна информации, которую доставляет нам ответ «да» или «нет» на некоторый вопрос, при условии, что мы заранее не имеем оснований считать один ответ более вероятным, чем другой.) В дальнейшем мы будем измерять информацию в «битах»: этому отвечает в формуле (*) значение $c = 3{,}32$ при обыкновенных десятичных логарифмах или $c = 1$ при двоичных логарифмах.

Предположим теперь, что наряду с опытом α мы имеем еще и другой опыт β, исходы которого также могут зависеть от случая (разумеется, опыт с заранее определенным исходом не может дать нам никаких данных о «случайном» опыте α). Возможно, что исходы опыта α никак не связаны с исходом β: так, например, не связаны между собой исходы опытов, состоящих в определении погоды на завтра и определении карты, вытянутой наугад из тщательно перемешанной колоды. (Это утверждение равносильно утверждению о невозможности предсказания погоды по картам.) В таком случае говорят, что опыты α и β независимы или что осуществление β не несет никакой информации об опыте α. Но часто может случиться так, что знание исхода β существенно влияет на вероятность того или другого исхода α: так, знание погоды на данный день существенно влияет на вероятность

[3] См. подробнее в кн.: А. М. Яглом, И. М. Яглом, Вероятность и информация, М., 1957, стр. 108—124.

той или иной погоды на следующий день; знание определенной буквы текста отражается на вероятностях появления отдельных букв после данной. Если разные исходы β влекут за собой разные изменения в распределении вероятностей для исходов опыта α, то опыты α и β называются *зависимыми*.

Итак, осуществление того или иного исхода может изменить вероятности $p_1, p_2, \ldots, p_n$ исходов другого опыта α, следовательно, и энтропию $H(\alpha)$. Предположим теперь, что мы многократно осуществляем опыт β; при этом каждый раз мы будем иметь свое значение величины $H(\alpha)$, зависящее от исхода β. *Среднее значение* всех этих величин обозначается через $H_\beta(\alpha)$ и называется средней условной энтропией α при условии предварительного осуществления β или просто *условной энтропией α при условии β*.

Ясно, что если опыты α и β независимы, то $H_\beta(\alpha) = H(\alpha)$. Если же опыты α и β зависимы, то $H_\beta(\alpha) < H(\alpha)$: предварительное осуществление опыта β может лишь уменьшить степень неопределенности опыта α. Разность $H(\alpha) - H_\beta(\alpha)$, показывающую, на сколько именно уменьшает неопределенность α предварительное осуществление β, Шеннон предложил назвать *информацией об опыте α, содержащейся в опыте β*, и обозначать через $I(\alpha, \beta)$.

Рассмотрим теперь третий опыт γ, предшествующий β. Каждому исходу опыта γ отвечает свое значение *условной энтропии* $H_\beta(\alpha)$; среднее значение всех этих значений $H_\beta(\alpha)$, отвечающее многократному повторению γ, мы назовем *условной энтропией второго порядка* опыта α (или условной энтропией α при условиях β и γ) и обозначим через $H_{\gamma\beta}(\alpha)$; при этом $H_{\gamma\beta}(\alpha) \leqslant H_\beta(\alpha)$. Аналогично этому, рассмотрев длинную цепочку опытов $\alpha_0, \alpha_1, \alpha_2, \alpha_3 \ldots$ (например, совокупность опытов по определению погоды в последовательные дни или по определению последовательных букв текста), мы можем сопоставить ей (убывающую) цепочку условных энтропий разных порядков $H(\alpha_0) \geqslant H_{\alpha_1}(\alpha_0) \geqslant H_{\alpha_2\alpha_1}(\alpha_0) \geqslant H_{\alpha_3\alpha_2\alpha_1}(\alpha_0) \geqslant \ldots$ Если наша цепочка опытов безгранична, то неограниченная последовательность убывающих положительных чисел $H_0 = H(\alpha_0)$, $H_1 = H_{\alpha_1}(\alpha_0)$, $H_2 = H_{\alpha_2\alpha_1}(\alpha_0) \ldots$ $\ldots, H_N = H_{\alpha_N\alpha_{N-1}\ldots\alpha_1}(\alpha_0), \ldots$ обязательно будет стремиться к некоторому предельному значению $H_\infty = H$. При $H = 0$ исход опыта α будет полностью определяться цепочкой (предшествующих ему) опытов $\alpha_1, \alpha_2, \alpha_3 \ldots$; если же $H \neq 0$, то даже и неограниченная цепочка опытов $\alpha_1, \alpha_2 \ldots$ не определит полностью исход α_0 (так, знание погоды за любое число предшествующих дней не определяет полностью знания завтрашней погоды).

Значение введенных понятий для общей теории связи в первую очередь определяется замечательной теоремой Шеннона, согласно которой *время, требующееся для передачи сообщения по любой линии связи (при условии, что передача ведется наиболее рациональным образом), прямо пропорционально количеству переданной информации*, измеренному в соответствии с формулой (*). Большое значение этих понятий для лингвистики определяется не только тем, что по техническим линиям связи чаще всего передается речь, но и тем, что, как показали широко поставленные психологические эксперименты[4], человек воспринимает информацию как линия связи с оптимальным режимом работы. При этом следует иметь в виду, что определенное, по Шеннону, «количество содержащейся в данном тексте информации» хотя и является в очень многих случаях весьма удачной характеристикой речи, не отражает все без исключения аспекты понятия информации. Существенным дефектом этого определения можно считать то, что оно опирается лишь на статистические характеристики речи и полностью игнорирует ее содержание; предпринятые в последнее время попытки дать, опираясь на идеи Шеннона, определение количества информации, учитывающее значимость содержания[5], пока еще нельзя считать полностью удавшимися.

Энтропия одной буквы текста. Избыточность языка

Теперь мы можем ответить на поставленный выше вопрос о «степени неопределенности» для моделей русского языка, иллюстрируемых «фразами» $\Phi_0, \Phi_1, \ldots, \Phi_4$, или для осмысленного текста.

Степень неопределенности опыта α_0, состоящего в выборе «наудачу» одной из n букв русского алфавита, равна

$$H_0 = H(\alpha_0) = -\frac{1}{n}\log\frac{1}{n} - \frac{1}{n}\log\frac{1}{n} \ldots - \frac{1}{n}\log\frac{1}{n} = -\log\frac{1}{n} = \log n,$$

ибо опыт α_0 имеет n (равновероятных) исходов, вероятность каждого из которых

[4] Обзор ряда таких экспериментов дан в заметке А. М. Яглома, в разделе «Новости математической науки», сб. «Математическое просвещение», вып. 5, М., 1959.

[5] См., например, Y. Bar-Hillel, R. Carnap, Semantic information, «The British journal for the philosophy of science», vol. IV, № 14, 1953.

равна $\frac{1}{n}$ (мы отбрасываем здесь множитель c в правой части формулы (*), считая, что основание системы логарифмов выбрано так, чтобы результат измерялся в нужных нам единицах, например в *битах*). Если условиться отождествлять между собой буквы *е* и *ё*, *ь* и *ъ*, но причислять к числу букв промежуток между словами («русский телеграфный алфавит», используемый в большинстве телеграфных кодов), то будем иметь $n = 32$ и, следовательно, $H_0 = \log_2 32 = 5$ *бит*; таким образом, определение одной взятой наугад буквы доставляет нам такую же информацию, как и пять ответов «да» или «нет». Можно считать, что число 5 *бит* указывает количество информации, содержащееся в одной букве «фразы» Φ_0.

Аналогично этому степень неопределенности опыта, состоящего в определении буквы «фразы» Φ_1, или информацию, содержащуюся в одной букве этой «фразы», можно приравнять энтропии $H_1 = H(\alpha_1) = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n$ опыта α_1, вероятности $p_1, p_2, \ldots, p_n$ исходов которого равны частотам букв русского алфавита. Точно так же информация, содержащаяся в одной букве «фразы» Φ_2, при составлении которой учитывались частоты всевозможных двубуквенных сочетаний, равна условной энтропии $H_2 = H_{\alpha_2}(\alpha_1)$ опыта α_1, состоящего в определении буквы текста при условии известности исхода опыта α_2 по определению предшествующей буквы; как отмечалось выше, $H_2 \leqslant H_1$. Информация, содержащаяся в одной букве «фраз» Φ_3 и Φ_4, равна условной энтропии второго и третьего порядка $H_3 = H_{\alpha_1\alpha_2}(\alpha_1)$ и $H_4 = H_{\alpha_1\alpha_2\alpha_3}(\alpha_1)$, где опыты α_3 и α_4 состоят в определении буквы текста, стоящей на два, соответственно, на три места раньше той буквы, выяснение которой составляет содержание опыта α_1. Точно так же можно определить условную энтропию $(N-1)$-го порядка $H_N = H_{\alpha_N\alpha_{N-1}\ldots\alpha_2}(\alpha_1)$ — условную энтропию того же опыта при условии предварительного определения $N-1$ предшествующих букв текста и, наконец, п р е д е л ь н у ю энтропию $H = H_\infty = \lim_{N\to\infty} H_N$, которой следует приравнять количество информации, содержащейся в одной букве осмысленного русского текста: $H_0 \geqslant H_1 \geqslant H_2 \geqslant H_3 \geqslant \ldots \geqslant H_N \geqslant \ldots \to H_\infty$.

Энтропия одной буквы текста (или «удельная энтропия») $H_\infty = H$ является самой важной статистической характеристикой языка. Однако часто удобнее характеризовать статистическую структуру языка, его внутренние закономерности не самим числом H_∞, а отношением $\frac{H_\infty}{H_0}$; равенство этого отношения единице означало бы полное отсутствие каких бы то ни было закономерностей, а равенство его нулю — то, что все буквы текста полностью определяются структурой языка и никакой произвол в них невозможен (разумеется, оба эти обстоятельства не имеют места ни для одного реального языка). Соответственно этому при сравнении статистической структуры различных языков, алфавиты которых имеют разное число букв, естественно сопоставлять между собой не значения энтропии H_∞, а значения отношения $\frac{H_\infty}{H_0}$.

Шеннон вместо этого рассматривал разность $R = 1 - \frac{H_\infty}{H_0}$, которую он назвал и з б ы т о ч н о с т ь ю языка. К сожалению, мы пока не имеем сколько-нибудь удовлетворительного определения энтропии H и избыточности R ни для одного языка; однако некоторые данные убеждают, что для всех европейских языков избыточность R заметно превосходит 50%.

Не совсем точно можно сказать, что избыточность R указывает процент «лишних» букв — если для какого-либо языка, скажем, $R = 60\%$, то это означает, что 60% букв являются лишними, т. е. могут быть восстановлены по остальным, исходя из внутренних закономерностей языка. Однако понимать это утверждение не следует слишком уж буквально. Разумеется, всякая возможность восстанавливать пропущенные или неправильно указанные буквы в осмысленном тексте (например, находить опечатки в книге или ошибки в телеграмме) основывается на существовании избыточности — в тексте типа «фразы» Φ_0, где избыточность равна нулю, никакая ошибка не может быть исправлена. (Соответственно этому в тех случаях, когда вероятность ошибки очень велика или ошибка очень нежелательна, мы намеренно повышаем избыточность, например повторяя каждое слово несколько раз или передавая текст «по буквам».) Однако точный смысл избыточности означает лишь, что некоторым специальным методом («кодированием») ровно $R\%$ (60% в нашем примере) букв можно «отбросить» (а часть оставшихся заменить другими), так что по оставшимся (и замененным) $(100 - R)\%$ букв можно будет восстановить безошибочно весь текст; при этом оставшиеся $(100 - R)\%$ букв будут образовывать «фразу» вроде Φ_0, между буквами которой не будет существовать уже никаких статистических связей [5а]. Далее можно утверждать, что если (в достаточно обширном

[5а] Специальные опыты, относящиеся к английскому языку, показали, что «случайным образом» в тексте можно отбросить не более 25% букв,— в противном случае нельзя будет догадаться, какие буквы были опущены.

или достаточно типичном тексте, в котором отражены общие статистические закономерности) отбросить больше $R\%$ букв, то их ни в коем случае нельзя будет восстановить. Так как в тексте, полученном отбрасыванием $R\%$ букв и заменой других, все буквы алфавита будут встречаться одинаково часто (ср. «фразу» Φ_0), то ясно, что в первую очередь целесообразно отбрасывать самые распространенные буквы, несущие, по Шеннону, наименьшую информацию. Поэтому наиболее безобидным является пропуск пробелов между словами (пробел есть самая частая «буква»), легко восстанавливаемый по оставшемуся тексту; далее, более частые гласные буквы несут меньшую информацию, чем согласные, с чем, по-видимому, связан пропуск этих букв в некоторых письменностях (семитские языки).

Конкретные данные об избыточности разных языков

		H_0	H_1	H_2	H_3
Английский язык	Письменная речь	4,76	4,03	3,32	3,10
Русский язык	Письменная речь	5,00	4,35	3,52	3,01
	Устная речь	5,38	4,77	3,62	0,70

Верхняя строка таблицы содержит данные статистической обработки обширного текста, приведенные в основополагающей для всего рассматриваемого направления работе Шеннона[6]. Соответствующие подсчеты для русского языка были произведены при помощи счетно-аналитических машин в Лаборатории систем передачи информации АН СССР Д.С.Лебедевым и В.А.Гармашем[7]; частоты отдельных букв и их сочетаний они получили, анализируя отрывки из «Войны и мира» Л. Н. Толстого объемом в 30 тыс. букв. Так как определение условных энтропий H_N высоких порядков требует весьма обширного статистического материала, то можно опасаться, что полученное этими авторами значение H_3 является не совсем точным; возможно, этим объясняется то, что в то время как H_0, H_1 и H_2 для русского языка превосходят соответствующую величину для английского языка (русский алфавит имеет 32 буквы против 27 латинского[8]), значение H_3 для русского языка, согласно Д. С. Лебедеву и В. А. Гармашу, оказывается меньшим, чем для английского.

Аналогичные данные для устной русской речи были получены американскими лингвистами Е. Черри, М. Халле и Р. Якобсоном[9]. Роль букв в их исследовании играли фонемы; общее число фонем они, следуя мнению ряда советских лингвистов, принимали равным 42. Основным источником, исходя из которого эти авторы определяли частоты отдельных фонем и их сочетаний, по-видимому, явился тщательно проведенный еще в 1925 г. А. М. Пешковским анализ 10 тыс. фонем разговорной русской речи[10]. Разумеется, столь малый объем анализируемого текста никак недостаточен для определения условных энтропий высокого порядка; поэтому хотя значение для H_1, приводимое Е. Черри, М. Халле и Р. Якобсоном, вероятно, довольно надежно, указанное ими значение H_3 не заслуживает доверия[11].

Рассмотрим еще таблицу, в которой приведены данные об энтропии H_1 для ряда европейских языков[12]: (напоминаем, что для всех языков, письменность которых использует латинский алфавит, $H_0 = \log_2 27 \approx 4{,}76$ *бит*):

---

[6] C. E. Shannon, Prediction and entropy of printed English, «Bell system technical journal», vol. 30, № 1, 1951.

[7] Д. С. Лебедев, В. А. Гармаш, Статистический анализ трехбуквенных сочетаний русского текста, сб. «Проблемы передачи информации», вып. 2, М., 1959.

[8] Здесь к числу букв причисляется пробел между словами.

[9] E. C. Cherry, M. Halle, R. Jakobson, Toward the logical description of languages in their phonemic aspect, «Language», vol. 29, № 1, 1953.

[10] См. А. М. Пешковский, Десять тысяч звуков, «Сборник статей», Л.— М., 1925.

[11] Вероятности отдельных фонем и их парных сочетаний были подсчитаны на большом современном материале в лаборатории фонетики Ленинградского университета. См. Л. Р. Зиндер, О лингвистической вероятности, сб. «Вопросы статистики речи (материалы совещания)», Л., 1958. Число фонем здесь принимается равным 48.

[12] G. A. Barnard, Statistical calculation of word entropies for four western languages, «IRE. Transactions on information theory», vol. I, № 1, 1955.

Язык	Немецкий	Французский	Испанский
H_1	4,10	3,96	3,98

Приближенное значение энтропии для слова $H_1^{(\text{слова})} = -P_1 \log P_1 - P_2 \log P_2 - \ldots - P_m \log P_m$ может быть вычислено, исходя из существующих частотных словарей языка или основываясь на так называемом законе Ципфа, согласно которому при упорядочении слов по их распространенности мы получим (*для любого языка*) следующие выражения для частот (вероятностей) P_1, $P_2 \ldots, P_m$ отдельных слов [13]: $P_1 = k$, $P_2 = \frac{k}{2}$, $P_3 = \frac{k}{3}$, ...; здесь k есть некоторая константа, определяемая экспериментально; общее число m слов естественно выбрать таким, чтобы было $P_1 + P_2 \ldots + P_m = k\left(1 + \frac{1}{2} \ldots + \frac{1}{m}\right) = 1$. Соответствующий подсчет, произведенный Шенноном [14], для английского языка (где $k = 0{,}1$), дал $H_1^{(\text{слова})} \approx \approx 11{,}82$ *бит*.

Разделив это выражение на $s + 1$, где s — среднее число букв слова (для английского языка $s = 4{,}5$), мы получим «энтропию, приходящуюся на одну букву». Ее можно считать несколько меньшей, чем $H_{s+1}^{(\text{буквы})}$, поскольку ясно, что статистические связи между $s + 1$ буквами одного слова являются более тесными, чем связи между $s + 1$ соседними буквами, принадлежащими разным словам. Таким путем Шеннон пришел к заключению, что для английского языка $H_7 \approx \frac{11{,}82}{5{,}5} \approx 2{,}14$ *бит*; отсюда следует, что для английского языка $R \geqslant 1 - \frac{2{,}14}{4{,}76} \approx 55\%$. Аналогичные подсчеты для четырех европейских языков (английского, немецкого, французского и испанского) произвел Г. Барнард [15], из результатов которого как будто следует, что избыточность испанского языка близка к избыточности английского, избыточность французского — ниже, а немецкого выше (все это нуждается еще в дальнейшей проверке) [16].

Другой путь определения избыточности для разных языков был также указан Шенноном [17], предложившим вместо определения частот (вероятностей) всевозможных N-буквенных сочетаний (число которых необычайно быстро растет с ростом N) рассматривать вероятности $q_i^{(N)}$ того, что N-ая буква текста будет правильно отгадана по известным $N-1$ предшествующим буквам с i-й попытки ($i = 1, 2 \ldots, n$, где n — число букв алфавита). При этом предполагается, что отгадывающий в совершенстве владеет статистическими закономерностями языка и делает предсказания наиболее рациональным образом. Такое владение не является чем-то исключительным. Дело в том, что каждый человек интуитивно владеет этими закономерностями для хорошо знакомого ему языка, так как ему приходится повседневно использовать эти знания, скажем, для восстановления букв и слов в неразборчиво сказанной фразе его собеседника. Все же из ряда отгадчиков желательно выбрать того, кто будет иметь лучшие результаты.

Ясно, что вероятности $q_i^{(N)}$ тесно связаны с вероятностями N-буквенных сочетаний; так, при $i = 1$ вероятности $q_1^{(1)}, q_2^{(1)} \ldots, q_n^{(1)}$, как легко понять, в точности совпадают с вероятностями $p_1, p_2, \ldots, p_n$ отдельных букв, расположенными в порядке убывания. Достоинством величин $q_i^{(N)}$ является то, что их можно определить экспериментально — при помощи опытов по отгадыванию букв, которые легко осуществить; следует только произвести достаточное число таких опытов (с тем, чтобы частоты соответствующих результатов можно было приравнять вероятностям $q_i^{(N)}$). Знание величин $q_1^{(N)}, q_2^{(N)}, \ldots, q_n^{(N)}$ дает возможность оценить значение условной энтропии $(N-1)$-го порядка H_N, ибо, как доказал Шеннон,

---

[13] G. K. Zipf, Human behavior and the principle of least effort, Cambridge (Mass.), 1949; ср. также: P. Giraud, Les caractères statistiques du vocabulaire, Paris, 1954.

[14] См. C. E. Shannon, Prediction and entropy...

[15] G. A. Barnard, указ. соч.

[16] Заметим, что намеченный здесь путь определения избыточности нельзя считать особенно надежным: напрашивающееся сопоставление величин $H_1^{(\text{слова})}$ и $H_{s+1}^{(\text{буквы})}$ трудно мотивировать достаточно убедительно.

[17] C. E. Shannon, Prediction and entropy...

$$2\,(q_2^{(N)} - q_3^{(N)})\log 2 + 3\,(q_3^{(N)} - q_4^{(N)})\log 3 + \ldots + (n-1)\,(q_{n-1}^{(N)} - q_n^{(N)})\log(n-1) + nq_n^{(N)}\log n \leqslant H_N \leqslant -q_1^{(N)}\log q_1^{(N)} - q_2^{(N)}\log q_2^{(N)}) - \ldots - q_n^{(N)}\log q_n^{(N)}. \quad (**)$$

Экспериментальное определение величин $q_i^{(N)}$ (а следовательно, и оценка условных энтропий H_N) было произведено для английского языка Шенноном [18] для $N = 1, 2, 3 \ldots$, 15 и $N = 100$; позднее эксперименты такого же типа на большем материале производились Н. Бертоном и Дж. Ликлайдером [19] для $N = 1, 2, 4, 8, 16, 32, 64, 128$ и $N \approx 10\,000$.

Наиболее интересный из полученных на этом пути результатов заключается в том, что, начиная, примерно, с $N = 30$, вероятности $q_i^{(N)}$ перестают зависеть от N, а это подсказывает, что $H_\infty \approx H_{30}$, т. е. что знание примерно 30 предшествующих букв текста дает почти полную возможную информацию о следующей за ними букве; любое число предшествующих букв сверх этих 30 уже почти не дает нового (но в то же время, скажем, H_{15} еще заметно больше H_∞). Что же касается до оценки величины $H = H_\infty$, а следовательно, и избыточности R, то на этом пути получается лишь довольно неточная оценка [при больших N левая и правая части двойного неравенства (**) не совпадают]. Наиболее широко поставленные эксперименты Бертона и Ликлайдера привели к заключению, что для английского языка R заключено между $^2/_3$ (т. е. 67%) и $^4/_5$ (80%); сходное значение избыточности получил и Шеннон. Такие же опыты для английской устной речи (опыты по отгадыванию фонем) произвел Д. Фрай [20], получивший, по-видимому, заниженное значение избыточности $R \approx 55\%$, что связано, быть может, с тем, что закономерности, относящиеся к фонемам, отгадчику менее привычны, и поэтому такие опыты требуют большой тренировки. Избыточность немецкого языка определял при помощи более упрощенной процедуры известный немецкий связист К. Кюпфмюллер [21], получивший значение избыточности, имеющее тот же порядок величины ($R \approx 70\%$ для письменной речи). Близкие результаты были получены для шведского языка Х. Хансоном [22], также использовавшим несколько упрощенный вариант метода Шеннона. К сожалению, неточность имеющихся методов не позволяет как следует оценить небольшие расхождения в избыточностях разных языков. Представляло бы бесспорный интерес более точное определение избыточности, позволяющее сравнивать ее для разных языков; в этой связи было бы желательно иметь возможно больше разных подходов к определению этой величины [22а].

Сравнение разных видов избыточности

В качестве основного элемента речи можно также рассматривать слова или фонемы (устная речь). При этом мы придем к новым значениям энтропии H_∞ и избыточности R.

Приведем здесь таблицу значений энтропии $H_1^{(\text{без проб.})}$, вычисленной для разных языков в предположении, что пробел между словами не включается в число букв (так что для русского языка теперь $H_0 = \log_2 31 \approx 4{,}95$ *бит* и для западных языков $H_0 = \log_2 26 \approx 4{,}70$ *бит*):

Язык	Русск.	Англ.	Нем.	Франц.	Исп.
$H_1^{(\text{без проб.})}$	4,46	4,14	4,095	3,98	4,01

Из сравнения этой таблицы с приведенными выше следует, что значение энтропии H_1 уменьшается, когда пробел между словами также включается в число букв;

[18] C. E. Shannon, Prediction and entropy...

[19] N. G. Burton and J. C. R. Licklider, Long-range constraints in the statistical structure of printed English, «American journal of psychology», vol. LXVIII, 4, Baltimore, 1955, стр. 650—653.

[20] D. B. Fry, Communication theory and linguistic theory, «IRE. Transactions on information theory», vol. I, 1953.

[21] K. Küpfmüller, Die Entropie der deutschen Sprache, «FTZ — Fernmeldetechnische Zeitschrift», Hf. 6, 1954.

[22] H. Hansson, The entropy of the Swedish language, «Proceedings of the 2-nd Symposium on information theory», Prague, 1960.

[22а] Возможно, что здесь может принести пользу анализ повторяемости отдельных букв и их сочетаний. Ср. Р. Л. Добрушин, Упрощенный метод экспериментальной оценки энтропии стационарной последовательности, «Теория вероятностей и ее применения», т. 3, № 4, М., 1958.

это связано с тем, что появление новой весьма вероятной «буквы» (вероятность пробела превосходит вероятность любой из остальных букв) делает менее неопределенным исход опыта по заданию одной «буквы» текста. Исключение в этом отношении составляет лишь немецкий язык, для которого характерна большая длина слов (средняя длина слова s составляет здесь 5,92 буквы против 4,5 букв в английском языке), в силу чего вероятность пробела здесь относительно мала и увеличение неопределенности опыта по одной букве текста, связанное с увеличением числа исходов, не компенсируется большей вероятностью этого исхода по сравнению с остальными.

Перейдем теперь к вопросу об избыточности $R^{(\text{без проб.})}$. Заметим, что величина H_∞ выражает, так сказать, «удельную информацию», или «информацию, которую несет одна буква текста»; большой отрывок из M букв несет «полную информацию» $H_\infty M$. Если записать один и тот же текст обычным образом и без пробелов между словами, то каждую из этих двух записей будет легко восстановить по другой: мы можем и отбросить все пробелы между словами, и восстановить пробелы в «сплошном», написанном без пробелов тексте на знакомом языке. Отсюда вытекает, что «полная информация», содержащаяся в обеих записях, будет одна и та же. Но число «букв» в тексте с пробелами будет в $\frac{s+1}{s}$ раз больше, чем число букв «сплошного» текста, где s — средняя длина слова (ибо один пробел приходится на s букв текста). Отсюда следует, что $H_\infty^{(\text{без проб.})} = H_\infty^{(\text{с проб.})} \cdot \frac{s+1}{s}$. Так как $H_0^{(\text{с проб.})} = \log n$, $H_0^{(\text{без проб.})} = \log(n-1)$ (где $n-1$ — число букв алфавита), то $\frac{H_\infty^{(\text{без проб.})}}{H_0^{(\text{без проб.})}} = \frac{H_\infty^{(\text{с проб.})}}{H_0^{(\text{с проб.})}} \cdot \frac{s+1}{s} \cdot \frac{\log n}{\log(n-1)}$, или $1 - R^{(\text{без проб.})} = (1 - R^{(\text{с проб.})}) \cdot \frac{(s+1)\log n}{s \log(n-1)}$. Следовательно, всегда $R^{(\text{без проб.})} < R^{(\text{с проб.})}$. В частности, для английского языка ($n = 27$, $s = 4{,}5$) имеем $1 - R^{(\text{без проб.})} = (1 - R^{(\text{с проб.})}) \frac{5{,}5 \log 27}{4{,}4 \log 26} \approx 1{,}29\,(1 - R^{(\text{с проб.})})$.

Те же соображения решают вопрос о сравнении избыточности для букв $R^{(\text{буквы})}$ и избыточности для слов $R^{(\text{слова})}$. Выше уже указывались методы вычисления $H_1^{(\text{слова})}$; но непосредственное вычисление «предельной энтропии» $H_\infty^{(\text{слова})}$ чрезвычайно усложняется как большим количеством слов, так и «дальнодействием» статистических связей между отдельными словами (появление в начале сколь угодно длинной книги слова «изопропилбензол» резко понижает вероятность встретить слово «морфема» в ее конце). Однако то обстоятельство, что знание всех букв текста равносильно знанию всех его слов, снова позволяет утверждать, что «полная информация», содержащаяся в некотором отрывке, может быть равным образом подсчитана по его словам и по буквам:

$$H_\infty^{(\text{слова})} \cdot \text{число слов} = H_\infty^{(\text{буквы})} \cdot \text{число букв}.$$

А так как число букв в среднем в $s+1$ раз превосходит число слов (здесь мы снова под $H_\infty^{(\text{буквы})}$ понимаем $H_\infty^{(\text{с проб.})}$, так что в среднем на одно слово приходится $s+1$ букв), то

$$H_\infty^{(\text{слова})} = (s+1) \cdot H_\infty^{(\text{буквы})},$$

или, если обозначить через m общее число слов языка, так что $H_0^{(\text{слова})} = \log m$, то

$$\frac{H_\infty^{(\text{слова})}}{H_0^{(\text{слова})}} = \frac{H_\infty^{(\text{буквы})}}{H_0^{(\text{буквы})}} \cdot (s+1) \cdot \frac{\log n}{\log m} \quad \text{или} \quad (1 - R^{(\text{слова})}) = (1 - R^{(\text{буквы})}) \cdot (s+1) \cdot \frac{\log n}{\log m}.$$

В частности, для английского языка, приняв $m = 100\,000$ [23], получаем $1 - R^{(\text{слова})} = (1 - R^{(\text{буквы})}) \cdot \frac{5{,}5 \log 27}{\log 100\,000} \approx 1{,}58\,(1 - R^{(\text{буквы})})$.

Наконец, те же соображения позволяют решить вопрос о связи избыточности устной речи с избыточностью речи письменной. Здесь мы снова исходим из возможности восстановить по написанному тексту все фонемы, а по устному — все буквы (т. е. из возможности прочесть письменный текст и записать устный); отсюда следует, что в том и в другом тексте содержится одинаковая «полная информация», т. е. $H_\infty^{(\text{фонемы})} \cdot$ число фонем $= H_\infty^{(\text{буквы})} \cdot$ число букв. Обозначим теперь через ω сред-

[23] Заметим, что наш результат лишь весьма слабо зависит от значения m.

нее число букв, приходящихся на одну фонему («среднюю длину фонемы»); величина ω есть важная статистическая характеристика языка, связывающая устную и письменную речь. В таком случае получим $H_\infty^{(\text{фонемы})} = \omega H_\infty^{(\text{буквы})}$. Следовательно, если k — общее число фонем, то

$$\frac{H_\infty^{(\text{фонемы})}}{H_0^{(\text{фонемы})}} = \frac{H_\infty^{(\text{буквы})}}{H_0^{(\text{буквы})}} \cdot \omega \frac{\log n}{\log k}, \text{ или}$$

$$1 - R^{(\text{фонемы})} = (1 - R^{(\text{буквы})}) \cdot \frac{\omega \log n}{\log k}.$$

В частности, отождествив для английского языка отдельные фонемы с фонетическими знаками, используемыми в наших англо-русских словарях, получим $k = 43$, $\omega \approx 1{,}2$ и, следовательно,

$$1 - R^{(\text{фонемы})} = (1 - R^{(\text{буквы})}) \cdot \frac{1{,}2 \log 27}{\log 43} \approx 1{,}05\,(1 - R^{(\text{буквы})}).$$

Сравнение разных текстов и разных языков

Все сказанное до сих пор относилось, так сказать, к «среднелитературному» языку. Однако представляет интерес изучение с рассмотренных здесь позиций и различных специальных текстов, например произведений определенных авторов или книг и статей по одной специальности. Ясно, что высокий уровень избыточности, скажем, в произведениях определенного автора, свидетельствует о злоупотреблении его одними и теми же словами и оборотами, т. е. о «плохой» литературной форме; напротив того, низкая избыточность в произведениях некоторых больших писателей может характеризовать яркость и нестандартность их языка. При этом следует иметь в виду, что слишком низкая избыточность может свидетельствовать также о специально усложненном и трудном для понимания языке: ведь нулевая избыточность характерна для хаотического набора букв вроде приведенной выше «фразы» Φ_0. Вообще все возможности совершенствования литературной формы произведений связаны с наличием в языке высокой избыточности; если бы речь (устная и письменная) представляла бы собой «оптимальную» систему кодирования, т. е. обладала бы нулевой избыточностью, то художественная литература, по-видимому, не была бы возможна. В этой связи возникает целый ряд интересных вопросов (например, о сравнении избыточности прозаической и поэтической речи), решение которых пока невозможно из-за бедности статистического материала.

Несколько более прост круг вопросов, связанный с изучением избыточности научно-технических текстов. Ясно, что любой научный жаргон (как и вообще каждый жаргон) имеет более высокую избыточность, чем «среднелитературный» язык, из-за относительной бедности его словаря и наличия ряда широко распространенных специальных терминов и выражений. Это заметно облегчает просмотр литературы по знакомой специальности и чтение такой литературы на малоизвестном языке. Высокая избыточность заметно облегчает также решение задачи о машинном переводе специальной литературы. Поэтому нам кажутся не особенно целесообразными тенденции к понижению избыточности научной литературы при помощи тщательно разработанной терминологии, как это имеет место, например, в деятельности авторитетного коллектива французских математиков, пишущих под общим псевдонимом «Никола Бурбаки». Впрочем, такого рода тенденции могут быть полезны как этап на пути к созданию «международных специальных языков»; создание таких языков, весьма облегчающееся высокой избыточностью специальных текстов, в настоящее время определенно стало на повестку дня (что также частично связано с вопросами механизации перевода и реферирования научных статей)[24]. Конкретных данных, касающихся избыточности отдельных специальных текстов, мы имеем пока очень мало. Можно упомянуть лишь об исследовании американских связистов Фрика и Самби, которые анализировали переговоры по радио между дежурным по аэропорту и пилотом; здесь избыточность может достичь 96% (меньшая избыточность в условиях высокого уровня шумов могла бы повлечь за собой катастрофу)[25].

Мало пока изученный вопрос о сравнении «веса» одного *бита* информации для разных языков заключается в следующем. При возможно более точном (без потери информации) переводе какого-либо текста на другой язык оба отрывка текста будут иметь одно и то же содержание, т. е. содержащаяся в них с м ы с л о в а я информа-

---

[24] Тенденции к созданию «международного языка», имеющего нулевую избыточность, привели также математиков к символическому «языку» современной математической логики.

[25] F. C. F r i c k, W. H. S u m b y, Control tower language, «The journal of the Acoustical society of America», vol. 24, № 6, 1952. Ср. также E. L. F r i t z, G. W. G r i e r, Pragmatic communication, сб. «Information theory in psychology», Illinois, 1955.

ция будет одна и та же. Если вычислить по нашим формулам содержащуюся в том и в другом тексте статистическую (шенноновскую) информацию, то при этом окажется, что *p бит* информации текста на одном языке эквивалентны *q битам* информации текста на другом языке, что позволит сравнить «смысловое содержание» 1 *бита* информации на том и на другом языке. Разумеется, необходимо, чтобы полученные таким путем выводы были проверены на достаточно обширном статистическом материале.

Исследования такого рода были проведены индийскими учеными Рамакришной и Субраманианом[26]. Они установили, что при переводе с английского языка на немецкий 1 *бит* содержащейся в английском тексте информации оказывается эквивалентным 1,22 *битам* информации, содержащейся в немецком тексте; это различие определяется как «экономностью» английского языка по сравнению с немецким, так и тем, что попытка совершенно точно передать содержание какого-либо отрывка на ином языке требует более распространенного по сравнению с подлинником изложения из-за трудностей, связанных с необходимостью сохранить в переводе все оттенки смысла, передаваемые иногда специфическими для данного языка средствами. С другой стороны, при переводе с немецкого на английский язык 1 *бит* содержащейся в исходном немецком тексте информации передавался 1,06 *битами* информации английского текста. Отсюда Рамакришна и Субраманиан заключили, что 1 *бит* информации, заключающейся в английском тексте, по смыслу равноценен примерно 1,15 *битам* информации немецкого текста; кроме того, по их мнению, процесс перевода требует еще дополнительной затраты информации (которую также можно оценить), связанной с необходимостью учета специфики другого языка. Эти, быть может, не вполне окончательные выводы бесспорно заслуживают внимания лингвистов.

Несмысловая информация

Устная речь, помимо «смысловой информации», несет еще весьма значительную несмысловую информацию, заключающуюся в индивидуальных особенностях голоса, в интонационных оттенках речи, в громкости речи; эта информация в отдельных случаях может даже противоречить смысловой (причем в таких случаях она, как правило, заслуживает большего доверия). Изучение несмысловой информации имеет большое практическое значение, так как передача этой информации по линиям связи (например, в случае вокальных передач) является очень важной технической задачей. Однако относящихся сюда точных данных известно пока очень немного.

Достаточно тщательно был изучен собственно только один довольно частный вопрос — о смысловых ударениях, выделяющих определенные слова фразы, и о заключающейся в них информации. Английский связист И. Берри на основании прослушивания «ряда типичных английских телефонных разговоров» указал частоты, с какими попадает под ударение то или иное слово[27]. Пусть $p_1, p_2, \ldots, p_m$ — вероятности (частоты) слов языка, расположенные в порядке убывания, а $S_1, S_2, \ldots, S_m$ — сами соответствующие слова; тогда, согласно наблюдениям Берри, вероятность того, что слово S_i (где i может быть равно 1, 2... или m) окажется под ударением, равна примерно $p_1 + p_2 \ldots + p_i$. Таким образом, редкие слова чаще выделяются ударением, чем частые. А так как вероятности $p_1, p_2, \ldots, p_m$ также можно считать известными, то не представляет труда вычислить среднее значение приходящейся на одно слово информации, зависящей от того, падает или не падает на слово ударение; эта информация оказывается близкой к 2,4 *бит/слово*[28]. Разумеется, возможно, что подмеченные Берри закономерности не сохраняют силу для других языков; возможно также, что и в английском языке они справедливы только для телефонных разговоров. Однако можно рассчитывать, что величина 2,4 *бит/слово* правильно указывает порядок величины информации, связанной с логическими ударениями.

Более широкий характер имеет исследование К. Кюпфмюллера[29], который использовал для оценки количества несмысловой информации, содержащейся в устной немецкой речи, данные физиологической акустики. Кюпфмюллер рассматривает всю несмысловую информацию, заключающуюся в устной речи, разбив ее на ряд групп и довольно грубо оценивая порядок величин; так, он всюду ищет лишь величину H_0, условно принимая, что избыточность во всех случаях равна 50%. Сравнивая количество смысловой и несмысловой информации, приходящееся на одну букву текста, Кюпфмюллер устанавливает, что при очень медленной речи несмысловая информация может составлять до 130% смысловой, а при очень быстрой — всего 30% смысловой; при нормальной скорости речи несмысловая информация составляет

---

[26] B. S. Ramakrishna, R. Subramanian, Relative efficiency of English and German languages for communication of semantic content, «IRE. Transactions on information theory», vol. 4, № 3, New York, 1958.

[27] I. Berry, Some statistical aspects of conversational speech, сб. «Communication theory», London, 1953.

[28] Ср. Б. Мандельброт, Закон Берри и определение «ударения», сб. «Теория передачи сообщений», М., 1957.

[29] K. Küpfmüller, указ. соч.

около 70 % смысловой информации. Различие в этих цифрах объясняется тем, что при быстрой речи нам гораздо труднее следить за интонационными оттенками речи; также и узнавание голоса становится при этом гораздо более трудным. По-видимому, подобное уменьшение количества несмысловой информации при увеличении скорости поступления смысловой информации связано с тем, что возможность человека через свои органы чувств воспринимать информацию и перерабатывать ее при помощи мозга является ограниченной[30].

Практические приложения

В заключение коснемся вопроса о практической ценности изложенных выше теорий и результатов. Нетрудно установить связь этих идей, скажем, с вопросами механического перевода или реферирования; представляют они интерес и для биологии и психологии (всестороннее изучение способности человека воспринимать и перерабатывать информацию); возможно, что кое-что здесь окажется полезным и для вопросов педагогики, связанных, например, с обучением иностранным языкам. Но наиболее непосредственное отношение теория информации имеет к задачам передачи речи по любым линиям связи. Методы кодирования речи даже для линии связи определенного типа могут быть очень разнообразными. Естественно искать такие коды, которые позволят вести передачу по возможности более экономно, например передавать за единицу времени возможно больше информации. Отыскание таких кодов весьма облегчается теоретико-информационными соображениями.

Уже самый старый из распространенных телеграфных кодов — код (или азбука) Морзе — учитывает статистические закономерности языка: более частые буквы имеют в нем более короткие обозначения, чем менее частые. Использование идей теории информации позволяет составить гораздо более экономные коды. В частности, телеграфные коды для индийских языков, которые разрабатывались лишь в последнее время, составлялись уже с полным учетом этих соображений[31].

Могут представить интерес для вопросов передачи сообщений и соображения о высокой избыточности «специальных» языков. В тех случаях, когда определенная линия связи служит для передачи в первую очередь каких-то сообщений специального характера, естественно при составлении кода для этой линии исходить из статистических данных, относящихся именно к этим сообщениям. Так, в настоящее время в США торговые коды больших фирм составляются с непременным использованием выводов теории информации и с учетом высокой избыточности, свойственной деловой корреспонденции фирмы.

Значение идей теории информации (и, в частности, их лингвистического аспекта) для теории связи в настоящее время все возрастает. Статистические методы уже начинают занимать важное место и при решении собственно лингвистических вопросов. Дальнейшее отставание в этой области советской лингвистики (в частности, отсутствие надежных данных о статистических характеристиках русского языка) совершенно недопустимо.

И. М. Яглом, Р. Л. Добрушин, А. М. Яглом

---

[30] Ср. с психологическими исследованиями, о которых говорилось выше.

[31] См. B. S. Ramakrishna, Information theory and some of its applications, «Current science», vol. 27, № 10, 1958.