

МАРК ВЕЙ

ЗАМЕЧАНИЯ ПО ПОВОДУ ЧАСТОТНОГО СЛОВАРЯ
ЧЕШСКОГО ЯЗЫКА

Превосходный частотный словарь чешского языка, подготовленный под руководством Я. Елинка, Й. Бечки и М. Тешителовой («Frekvence slov, slovních druhů a tvarů v českém jazyce», в печати), содержит статистические сведения, позволяющие проверить на чешском материале данные о статистической структуре словаря, установленные ранее для других языков. Они позволяют также получить и новые числовые данные.

1. В статье о статистике чешского словаря¹ я показал соотношение между количеством различных слов, употребленных одновременно в определенном количестве текстов, и количеством этих текстов. Это соотношение, само по себе представляющее интерес для статистики, интересно еще и тем, что в нем в форме математической зависимости утверждается, что для получения основных сведений о статистической структуре словаря языка не обязательно неограниченно увеличивать выборку.

Вместе с тем, если построить график для указанной зависимости, окажется, что полученная кривая в общем совпадает с кривой, которую можно было бы получить экспериментальным путем на основании прямого подсчета. Исключение составят две крайние области — явление обычное в статистике. Это побуждает дать лингвистическую оценку результатов.

а) Как бы ни было мало количество различных слов, каждое из которых встретилось во всех 75 текстах (рассмотрение которых легло в основу словаря «Frekvence slov»), оно оказывается больше, чем позволило бы ожидать констатированное отношение. Это связано с тем, что в число таких слов попадают не столько слова чешского языка с наивысшей частотностью, сколько слова, без которых невозможно никакое высказывание любой протяженности. Указанное количество не может бесконечно уменьшаться.

б) Количество слов, зарегистрированных в очень небольшом количестве текстов (что обозначается графически другой крайней областью кривой), равным образом оказывается больше, чем можно было бы ожидать. Объясняется это тем, что на этом уровне появляется значительное количество словарных единиц, не принадлежащих к узаконенному словарю языка (в особенности имена собственные и слова им подобные).

2. В моей статье в «Slavia»² при рассмотрении главных числовых зависимостей, ранее установленных статистиками для других языков, я обратился к проверке их на чешском материале. Для большинства случаев справедливость этих зависимостей подтверждается. В особенности примечательно, что для зависимости между вероятностью употребления и зна-

¹ M. V e y, A propos de la statistique du vocabulaire tchèque, RESL, t. XXXIV, fasc. 1—4, 1957.

² M. V e y, A propos de la statistique du vocabulaire tchèque. Examen de principales relations numériques, «Slavia», rocn. XXVII, seš. 3, 1958.

чимостью слова (последняя пропорциональна числу фонем) я получил формулу, идентичную той, которая иным путем уже была получена П. Гиро.

Однако необходимо отметить одно серьезное расхождение. Согласно Г. К. Зипфу, среднее число значений данного слова прямо пропорционально квадратному корню из его частоты. В чешском же языке среднее число значений прямо пропорционально корню четвертой степени из частоты. В моей статье в «Slavia» я ограничился констатацией этого факта, поскольку задачей статьи не было сравнение данных, полученных для чешского языка, с данными в особенности важными для какого-нибудь другого языка, например для французского. Однако такое сравнение не лишено интереса. С этой целью следовало бы обратиться не к абсолютным частотам, а к вероятностям, поскольку протяженность текстов, подвергнутых обследованию для составления «Frekvence slov» (более 1623 тыс. слов) и «French word book» В. А. С. Хенмона (400 тыс. слов), различна. Не входя в подробности вычисления, заметим, что слова, вероятность

которых равна $\frac{26}{1000000}$, имеют в чешском и во французском языках одно и то же среднее число значений (а именно 3). У слов с низшей вероятностью (это слова редкие) больше значений в чешском языке (от трех до одного с небольшим), чем во французском. Напротив, слова употребительные, с вероятностью больше чем $\frac{26}{1000000}$, менее многозначны в чешском по сравнению с французским. Это весьма показательное различие является результатом различия в структурах чешского и французского словарей: у чешских слов, более многочисленных, количество значений в общем меньше, чем у французских слов.

В этой связи следует обратить внимание на следующее весьма важное обстоятельство. Обычно принимают за истину, верную в отношении всех языков, что некоторое небольшое число слов составляет высокий процент от всех слов, заключенных в тексте. Считается, что 15 наиболее употребительных слов всегда составляют приблизительно $\frac{1}{4}$ часть от общего количества употребленных слов, 66 наиболее употребительных слов — примерно половину, 320 — около $\frac{3}{4}$, 1100 — около $\frac{9}{10}$. Эти цифры могут оказаться верными для различных языков, например для немецкого, однако они должны быть полностью пересмотрены в отношении чешского языка (а также, возможно, и в отношении других славянских языков). В чешском языке необходимо знать 21 единицу словаря для понимания $\frac{1}{4}$ части всех слов, содержащихся в тексте, 240 — для понимания половины, более 2 тыс. — для понимания $\frac{3}{4}$, 6 тыс. — для 87%, 8 тыс. — для $\frac{9}{10}$.

Перевели с французского О. А. Лантева и М. М. Маковский