ПСИХОФИЗИЧЕСКИЕ ОСНОВЫ ЗАВИСИМОСТИ “УПОТРЕБИТЕЛЬНОСТЬ – ПОЛИСЕМИЯ”

В.В. Кромер

Новосибирский государственный педагогический университет
630126, Новосибирск-126, Вилюйская 28
e-mail: applied@nspu.nsu.ru

Известно множество исследований, посвященных связи между употребительностью слова и его полисемичностью. Как правило, за меру полисемичности принимается количество значений слова, определяемое по одноязычному толковому словарю (ТС). В качестве меры употребительности принимается частота слова в достаточно представительной выборке (корпусе текстов), либо его ранг в частотном словаре (ЧС), составленном по исследуемой выборке. Сопоставление двух рядов данных позволяет подобрать эмпирическую формулу, выражающую количество значений в функции частоты либо ранга.

Исследователями предлагался ряд формул, выражающих рассматриваемую зависимость. Для описания зависимости предлагалась степенная функция (Ю.А. Тулдава):

$$m = \alpha F^{\gamma}, \qquad (1)$$

где m – количество значений, F – частота, α и γ – параметры зависимости.

Подобная же зависимость описывалась логарифмической функцией (М.В. Арапов):

$$m = \{D - a \ln i \quad : i < i_0; \quad 1 \quad : i > i_0\}, \qquad (2)$$

где i – ранг слова, i_0 – граничное значение ранга, начиная с которого $m = 1$, D и a – параметры зависимости.

Параметры в формулах (1) и (2) определяются исходя из наилучшего приближения эмпирических данных в некотором интервале аппроксимации. Основной недостаток формулы (1) – невозможность экстраполяции за пределы интервала аппроксимации. В формуле (2) подобный недостаток устраняется кусочным определением зависимости.

В настоящей работе использован иной подход. Постулируются некоторые достаточно общие положения и на экспериментальном материале проверяются следствия из них. Для исследования выбран наиболее представительный ЧС русского языка п/р Л.Н. Засориной [1]. В качестве ТС желательно использование гаммы словарей со следующими свойствами:

1. Объемы словников ТС последовательно возрастают в пределах гаммы.

2. В основе ЧС и всех ТС лежит одна и та же аналитическая грамматика.

3. Все ТС являются вложенными один в другой, т.е. весь словник меньшего по объему словаря (“младшего”) входит в больший словарь (“старший”).

4. Количество значений конкретного слова в младшем словаре не превышает количества значений его в старшем словаре, т.е. меньший по объему словарь создается на основе большего путем отсечения части значений.

5. Количество слов в ЧС не менее количества слов в самом богатом ТС, а словники ТС образуются из словника ЧС путем отсечения соответствующего количества слов с конца рангового распределения.

Рассмотренные требования являются идеализированными, и, насколько нам известно, подобной гаммы толковых словарей, согласованных с ЧС, для русского языка не существует. Нами использован ряд ТС русского языка, созданных в разное время отдельными авторами и коллективами составителей [2, 3, 4, 5]. Названия словарей и их характеристики приведены в таблице 1.

Таблица 1

№	Название толкового словаря	Краткое обозначение	Характеристика	Объем словника
1	Словарь современного русского литературного языка	ССРЛЯ	Большой толковый словарь	120000
2	Словарь русского языка в четырех томах	СРЯ	Средний толковый словарь	83000
3	Ожегов С.И. Словарь русского языка	СО	Краткий толковый словарь популярного типа	57000
4	Краткий толковый словарь русского языка	КТС	Учебный толковый словарь для нерусских учащихся	5000

Положения, лежащие в основе исследования:

1. Каждое употребление слова дает новое его значение, т.е. количество значений слова равно общему количеству зафиксированных его употреблений.

2. Каждый язык (подъязык, идиолект) базируется на определенном корпусе текстов и вследствие этого ограничен.

3. Носитель конкретного языка в общем случае не различает единичные значения слова, связанные с отдельным словоупотреблением. Значения группируются им в группы значений, при этом самая малая группа включает одно значение, основанное на одном употреблении слова.

4. Каждый ТС базируется на определенной выборке (корпусе текстов). Данную выборку предлагается назвать конститутивной.

5. Количество групп значений определяется в соответствии с психофизическим законом Вебера-Фехнера и асимптотически стремится к $(\ln F + C)$, где F – частота (количество употреблений) слова в корпусе текстов, положенном в основу подъязыка, а C = 0,5772... – постоянная Эйлера. Более точно математическое ожидание количества групп значений определяется через пси-функцию:

$$m = \psi(F + 1) + C\,. \tag{3}$$

6. ТС также толкуют лишь группы значений, при этом множество толкуемых групп значений (традиционно называемых количеством словарных значений) является составным множеством и определяется операциями над нечеткими множествами групп значений данного слова в отдельных подъязыках рассматриваемого языка.

Сделаем предположение, что корпусу текстов, на котором базируется ТС, свойственно ципфовское распределение частот слов:

$$F = \frac{K}{i^{\gamma}}, \tag{4}$$

где F – частота, i – ранг, K и γ – параметры.

Количество словарных значений в соответствии с принятой моделью составляет:

$$m = \psi(F + 1) + C \approx \ln F + C = \ln K - \gamma \ln i + C\,. \tag{5}$$

Обозначив $(\ln K + C)$ за D, а γ за a, получаем известную зависимость (2).

В работе [6, с. 19] показано, что горизонтальное распределение пси-функции частоты симметрично и близко к нормальному, что позволяет использовать для определения параметров K и γ в формуле (4) соотношения:

$$\begin{cases} \sum_{i=1}^{v} m = \sum_{i=1}^{v} \left[\psi\left(\frac{K}{i^{\gamma}} + 1 \right) + C \right], \\ \frac{K}{v^{\gamma}} = 1 \end{cases} \tag{6}$$

где m – количество значений слова в ТС, v – объем словника ТС. Второе выражение в (6) требует равенства частоты самого редкого слова конститутивной выборки (и соответственно количества его значений в ТС, поскольку $\psi(2) + C = 1$), единице.

Предварительное исследование показало, что одновременно условиям (3) и (4) во всем диапазоне рангов ЧС [1] удовлетворяет лишь словник СО. У меньших и больших ТС наблюдаются регулярные (т.е. зависящие от объема словника ТС) отклонения в области рангов 1–100, поэтому при определении K и γ по системе (6) нижний предел суммирования в верхней формуле принят равным 101. Верхний предел суммирования для словарей [2, 3, 4] берется равным количеству зафиксированных в ЧС [1] слов, т.е.

около 40000. На этапе предварительного исследования суммирование производилось по прореженной сетке рангов.

Вычисленные значения K и γ по четырем ТС (таблица 1) позволяют построить зависимость $m_t(i)$, где m_t – теоретическое значение количества словарных значений в зависимости от ранга слова. Отмечены следующие закономерности:

1. Значения m_t для слова с рангом 1 возрастают с ростом v.

2. Значения γ возрастают с ростом v.

3. Сглаженная эмпирическая зависимость $m(i)$ достаточно хорошо описывается теоретической $m_t(i)$.

Зависимости $m_t(i)$ и $m(i)$ наносятся на график с логарифмическим масштабом по оси рангов и линейным по оси количества значений. При выбранных осях зависимости практически линейны. График в билогарифмических осях не имеет лингвистического смысла, поскольку величины m_t и m уже являются логарифмическими по своей психофизической природе, и подобного рода зависимости в билогарифмических осях отображаются выпуклой кривой.

Для исследованных ТС значения γ находятся в пределах 0,3–0,7. В работе [6, с. 20] распределения с $\gamma < 1$ рассматриваются как суперпозиция 2 распределений: первообразного ципфовского распределения с близким к 1 значением γ и зависимости, отражающей особенности горизонтального распределения. Низкое значение γ свидетельствует о возрастании дисперсии горизонтального распределения пси-функции частоты с ростом ранга, т.е. положенные в основу ТС выборки являются резко неоднородными, и степень неоднородности выше для выборок, положенных в основу меньших словарей.

Объем конститутивной выборки N может быть оценен на основе теоретической зависимости F по формуле:

$$N = \sum_{i=1}^{v} F = \sum_{i=1}^{v} \frac{K}{i^{\gamma}}. \qquad (7)$$

Типичное значение N составляет от нескольких десятков тысяч словоупотреблений (для КТС) до нескольких сотен тысяч словоупотреблений (для ССРЛЯ).

Операции над нечеткими множествами, элементами которых являются значения слова в отдельных подъязыках, определяют количество значений рассматриваемого слова, толкуемых в конкретном ТС. При этом при переходе к меньшему ТС происходит уменьшение математического ожидания количества значений слова, характеризуемое некоторым коэффициентом пропорциональности, в первом приближении инвариантным к рангу слова в ЧС.

Рассматриваемая модель позволяет также прогнозировать распределение частот отдельных значений слова (в ЧС [1] аспект семантического варьирования и эквивалентности не принимался во внимание). Распределение ранжированных по

убыванию вероятностей отдельных значений слова асимптотически стремится к геометрическому распределению с параметром $q = \frac{1}{e} \approx 0{,}368$:

$$p_n = (1-q)q^{n-1} = \frac{e-1}{e^n}. \qquad (8)$$

где p_n – вероятность n-ого значения. При достаточном количестве значений вероятность первого значения слова составляет 0,632, второго – 0,233, третьего 0,086 и т.д.

Литература

1. Частотный словарь русского языка / Под. ред. Л.Н. Засориной – М.: Рус. яз., 1977.

2. Словарь современного русского литературного языка: В 17 т. – М–Л.: Изд-во АН СССР, 1948–1965.

3. Словарь русского языка: В 4 т. – М.: Рус. яз., 1985–1988.

4. Ожегов С.И. Словарь русского языка. – М.: Рус. яз., 1978.

5. Краткий толковый словарь русского языка. – М.: Рус. яз., 1990.

6. Кромер В.В. Подпорно-экспоненциальная модель генеральной лексической совокупности английского языка / Новосиб. гос. пед. ун-т. – Новосибирск, 1997. – Деп. в ИНИОН РАН 18.12.97, № 53134.