Р. М. ФРУМКИНА

СТАТИСТИЧЕСКАЯ СТРУКТУРА ЛЕКСИКИ ПУШКИНА *

При изучении языка писателя мы, как правило, не можем точно указать, в чем заключается «богатство», специфика лексики данного автора, хотя, несомненно, ощущаем эти качества интуитивно. В науке неоднократно делались попытки подойти к вопросам изучения лексического фонда писателя на основе количественного критерия — частоты употребления тех или иных слов в его произведениях¹. Количественным критерием предлагалось также воспользоваться в целях эвристики². На этом пути есть значительные успехи, есть и неудачи.

Материалы «Словаря языка Пушкина»³, где для каждого слова указано, сколько раз оно употреблено в пушкинских текстах, впервые позволяют предпринять такого рода исследование для русского автора. Даже предварительные результаты, полученные при таком подходе, представляют, как нам кажется, существенный интерес.

Еще в сороковых годах XX в. было достоверно установлено, что лексика каждого достаточно длинного текста, будь то художественная или научная литература, обладает определенной статистической структурой⁴. В общих чертах это означает, что для каждого автора существует определенное строгое соотношение в употреблении более частых, менее частых и редких слов и что в зависимости от того, каково это соотношение, мы субъективно ощущаем словарь автора как богатый, разнообразный, или бедный, однообразный. Соотношение такого рода изучалось для английских, французских, латинских писателей; для русских авторов таких данных в литературе не имеется.

При определении такого соотношения для лексики Пушкина нами были получены следующие данные. Общий объем пушкинских текстов (по академическому изданию без черновых вариантов) — 544 777 слов. Всего разных слов у Пушкина (по данным картотеки словаря Пушкина и в соответствии с тем пониманием слова, которое принято в словаре) — 21 197. При этом 50 самых частых слов составляют 38% всех словоупотреблений, т. е. более трети всех пушкинских текстов. 200 самых частых слов составляют 52%, т. е. более половины текстов. 1000 самых частых слов составляют 70% текста: это значит, что для лица, не владеющего русским языком, достаточно выучить эти 1000 слов, чтобы из каждых 10 000 слов текста понимать 7000 слов без обращения к словарю. Если мы продолжим это рассмотрение, то окажется, что 8000 самых частых слов соответствуют 95% текста Пушкина, а на остальные 13 200 слов пушкинской лексики

* В настоящей статье излагаются первые результаты совместной работы по статистическому исследованию лексики Пушкина группы «Словаря языка Пушкина» Института русского языка АН СССР и сектора прикладного языкознания Института языкознания АН СССР.

¹ См. G. U. Yule, *The statistical study of literary vocabulary*, London, 1944; M. L. Hanley, *Word index to James Joyce's Ulysses*, Madison, 1951 [mimeogr. изд.]; W. Kuraszkiewicz, *Statystyczne badanie słownictwa polskich tekstów XVI wieku*, «Z polskich studiów slawistycznych. Prace język. i etnogen...», Warszawa, 1958.

² См. В. В. Виноградов, *Лингвистические основы научной критики текста*, ВЯ, 1958, №№ 2 и 3.

³ «Словарь языка Пушкина», М., т. I—1956, т. II—1957, т. III—1959 (т. IV — рукопись).

⁴ См. G. U. Yule, указ. соч.

приходится только 5% текста. Очевидно, что если 13 200 слов (т. е. 62% лексики) составляют всего 5% текста, то мы имеем дело с весьма редкими словами.

Рассмотрим эту группу из 13 200 слов (которые мы в дальнейшем будем условно называть «редкими») с точки зрения того, как часто встречаются в текстах Пушкина входящие в нее слова. Получаем следующие результаты: 1) около 6400 слов (6389) употреблены Пушкиным всего по одному разу, 2) около 3000 слов (2913) — по два раза, 3) самые частые слова из указанных 13 200 встречаются во всех текстах Пушкина не более пяти раз.

Таким образом, слова, встречающиеся у Пушкина не более двух раз, составляют 48% (почти половину!) его лексики. Если мы сочтем, что слова, употребленные у Пушкина по одному разу, не существенны для характеристики его словаря, и будем рассматривать лексику Пушкина без них, то мы обедним пушкинский словарь на 30%, а если мы оставим без внимания все слова, употребленные не более двух раз, то лексика Пушкина уменьшится вдвое. Более того, если бы мы по каким-либо соображениям условились рассматривать только, так сказать, «употребительные» слова, например встретившиеся на протяжении всех томов академического издания не менее 25 раз, то полученный таким образом словарь свелся бы к небольшой группе в 3891 слово. Из этих данных с очевидностью следует вывод о том, что богатство словаря автора заключено по преимуществу в редких словах: именно за их счет мы имеем словарь в 21 197 слов. Таким образом, слова, встречающиеся у автора один раз, должны рассматриваться как существенная часть его лексики.

Продолжим рассмотрение группы редких слов с точки зрения качественной. Ограничимся пока словами, встречающимися по одному разу, — их у Пушкина 6389. В целом об этой группе слов можно сказать следующее: 1) большинство этих слов стилистически нейтрально; 2) наряду с большим количеством окказиональных слов — прилагательных, образованных от имен собственных и географических названий, прозвищ и т. п., в разряд «редких» попадают и такие слова, которые интуитивно мы считаем естественно входящими в пушкинскую лексику. Так, всего по одному разу встречаются у Пушкина прилагательные *благодарный*, *безграничный*, *береговой*, *бравый*.

На основании этого можно высказать одно соображение, небезынтересное для эвристики. Предположим, рассматривается некоторый текст, принадлежность которого Пушкину оспаривается. В качестве одного из аргументов против атрибуции текста Пушкину приводится тот факт, что в тексте есть слова, не засвидетельствованные в картотеке словаря Пушкина и не встречающиеся в черновых вариантах. Однако, если уже упоминавшееся выше слово *благодарный* употреблено Пушкиным только в стихотворении «В часы забав иль праздной скуки», то для отсутствия этого слова в словнике Пушкина достаточно было, чтобы это стихотворение временно оставалось по каким-либо совершенно случайным внешним причинам неизвестным. Если же имеют место попытки атрибуции довольно длинного текста, например длиной в 500 слов и более (что равно приблизительно 2 страницам прозаического текста), то в таком тексте может быть целая группа слов, в других текстах предполагаемого автора не встречающихся, причем это могут быть и не существенные, которые более всего связаны с темой произведения, а прилагательные, наречия и глаголы. Например, у Пушкина всего один раз встречаются такие наречия, как *бедно*, *беспокойно*, *беспоощдно*, *бесчестно*, *вдохновенно*, *горделиво*, *горестно*, *завистливо*, *заманчиво*. Вообще надо сказать, что в тех случаях, когда количественный критерий применялся успешно в целях эвристики, исследователям потребовалось составить целую совокупность весьма сложных статистических характеристик⁵.

⁵ См., например, G. U. Yule, указ. соч.

Несколько слов о жанровой характеристике так называемых «редких слов». Если разделить все тексты Пушкина на четыре жанра: 1) стихотворные тексты, 2) художественная проза, 3) публицистика и исторические сочинения, 4) письма — и сгруппировать слова, встречающиеся не более одного раза, по жанрам, то окажется, что для каждого такого слова (если отвлечься от имен собственных мифологического характера) трудно объяснить причину его употребления именно в данном жанре. В особенности это касается глаголов, прилагательных и наречий (так, единственное употребление слова *златовласый* приходится на художественную прозу). Однако, рассматривая в целом группу слов, употребленных только в данном жанре, мы все же можем выделить лексические пласты, характерные, например, для публицистики и исторических сочинений Пушкина в противоположность художественной прозе. Можно предположить, что лучше всего различие в употреблении слов в разных жанрах прослеживается на словах средней частоты, встретившихся во всех текстах Пушкина примерно от 10 до 50 раз. Это подтверждается предварительными результатами.

Для примера рассмотрим наугад несколько слов, употребленных по 30 раз. При этом выявляются слова, относительно равномерно распределенные по жанрам, например *великолепный*, и слова, сосредоточенные в отдельных жанрах. Преимущественно в стихах употребляются слова: *богиня* (30 из 30), *воинственный* (21 из 30), *буйный* (24 из 30), *небрежный* (24 из 30), *дороже* (21 из 30). Совсем не встречаются в стихах слова *императорский*, *мелочной*. По одному разу из 30 встречаются в стихах слова: *занимательный*, *деревянный*, *иностранец*. Нам кажется, что, продолжая работу в этом направлении, мы можем получить интересные результаты, характеризующие лексическую структуру, присущую тем или иным жанрам пушкинского творчества.

Перейдем теперь к анализу группы самых частых слов у Пушкина. Если расположить весь словник в порядке убывания частот, от самых частых слов к самым редким, и рассмотреть первые 200 слов в этом списке, т. е. 200 самых частых слов у Пушкина, то выясняется следующее: 1) самое частое из них — *и* — употреблено у Пушкина 25 026 раз, самые редкие — *случай* и *чувство* — по 297 раз каждое (отметим предварительно, что 300 раз на более чем полмиллиона слов пушкинских текстов — это не так уж часто); 2) среди этих 200 слов нет ни одного стилистически нейтрального; 3) подавляющее большинство этих слов и сейчас являются самыми частыми для русского литературного языка. Во всяком случае резкие сдвиги в частоте употребления можно отметить только для 20 слов из 200. Это следующие группы слов: а) служебные слова, местоимения и наречия, которые употребляются в современном языке или крайне редко или с определенной стилистической целью: *сей*, *да* (союз), *пред*, *хоть*, *кой*; б) слова, ставшие значительно менее употребительными в силу изменения общественного строя и исторической обстановки: *бог*, *государь*, *господин*, *князь*, *граф*, *царь* (всего 6 слов); в) слова, ставшие менее употребительными в результате изменения системы значений или стилистической характеристики: *душа* (ср. *сердце*), *стих* (ср. *стихи*, *стихотворение*), *милостивый* (в обращении *милостивый государь* устарело), *взор* (у Пушкина это слово нейтрально, но ср. современное противопоставление *взор* — *взгляд*), *войско* (ср. *войска*, *армия*), *должный*, *уж* (в значении современного *уже*), *милый* (ср. *дорогой*).

Еще более показательным является такое сравнение для 500 самых частых слов (мы продолжаем наш список, добавляя 300 следующих по частоте). Теперь самые редкие слова из рассматриваемых имеют частоту 143 (это слова: *воображение*, *глава*, *гроб*, *кругом*, *легкий*, *луна*, *поехать*, *следующий* и *слишком*). Значительные сдвиги в употребительности отмечаем для следующих слов: а) служебные слова, местоимения и наречия, которые для современного языка или архаичны, или стилистически окра-

шены: *оний, ибо, противу, друг* (краткая форма от *другой*), *меж, было* (частица), *тотчас, ныне, столь, ужели* (всего 10 слов); б) слова, ставшие значительно менее употребительными в силу общественно-исторических изменений: *царский, величество, государыня, графиня, слуга, дама, бал* (всего 7 слов); в) слова, ставшие менее употребительными в результате семантико-стилистических сдвигов: *дева, внимать, дух, младой, глава, любезный, муза, надобно, должно, батюшка, словесность* (всего 11 слов). Итак, для 500 самых частых слов существенные изменения в сторону уменьшения частоты употребления отмечены для 48 слов.

Остальные слова представляются нам вполне современными и почти без исключения стилистически нейтральными. Кроме того, они совершенно не содержат устаревших словообразовательных элементов. Например, с № 290 по № 300: *вероятно, писатель, ужасный, род, скоро, шум, звать, помнить, дочь, важный, разговор*; с № 340 по № 349: *богаться, отправиться, иной, нежный, темный, вместе, внимание, петь, вновь, возвратиться*.

Это позволяет высказать следующее предположение. Наше субъективное ощущение языка Пушкина как очень близкого к современному, несомненно, связано с тем, что 500 самых частых слов⁶ его словаря включают всего 48 сколько-нибудь устаревших или значительно изменивших систему значений, поскольку эти 500 слов соответствуют примерно 62% пушкинского текста.

Полученные результаты позволяют поставить следующий вопрос. Несомненно, что, например, язык Карамзина является по отношению к современному состоянию русского литературного языка более ранней ступенью, чем язык Пушкина. Вместе с тем в сочинениях Карамзина, несомненно, встречаются все или почти все те слова, которые входят в 500 самых частых слов у Пушкина. Будут ли эти слова у Карамзина входить в ту же частотную группу, что и у Пушкина? В качестве рабочей гипотезы можно высказать мысль о том, что списки «500 самых частых слов у Пушкина» и «500 самых частых слов у Карамзина» совпадут лишь частично: часть тех слов, которые у Карамзина принадлежат к 500 самым частым словам, у Пушкина окажутся употребляемыми реже и войдут, например, лишь в список «2000 самых частых слов», и т. д. Представляется, вообще говоря, возможным, располагая лексику каждого из двух или более авторов в порядке убывания частот и рассматривая с разных точек зрения степень коррелированности таких списков, найти некоторые объективные характеристики словаря именно данного писателя. Можно полагать также, что исследование статистических характеристик словарей различных писателей даст возможность выделить индивидуальные особенности лексики писателя от фона, характерного для эпохи в целом.

⁶ Цифра «500», конечно, условна: можно было бы рассматривать группу в 600, 700, 1000 самых частых слов.