ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 4 1983

КРИТИКА И БИБЛИОГРАФИЯ

ОБЗОРЫ

АЛЕКСЕЕВ П. М.

О НОВЫХ РАБОТАХ В СЛАВЯНСКОЙ СТАТИСТИЧЕСКОЙ ЛЕКСИКОГРАФИИ

Становится все более очевидным, что нецелесообразно ограничивать сферу статистической лексикографии словарями исключительно частотными. Статистические словари других видов составляются и публикуются; они привлекают внимание специалистов широкого круга; предпринята попытка их классификации [1]. Однако частотные словари (ЧС) по-прежнему преобладают в статистической лексикографии [2], а появление очередного ЧС становится заметным событием не только для лингвостатистики.

За последние годы опубликованы два первых выпуска из серии ЧС белорусского языка [3], серия из пяти ЧС польского языка [4], совсем недавно вышел в свет двухтомный ЧС, открывший серию ЧС украинского языка [5]. Эти серии в большей или меньшей степени различаются по оформлению входящих в них ЧС, по способу составления, по объему и отбору текстового материала, по принципам анализа лингвистических единиц текста.

Общее же для них заключается не только в том, что они отражают первый опыт выявления лексики, наиболее употребительной в этих языках. Они убедительно демонстрируют, что составление ЧС, представляющего письменную форму «языка в целом»,— работа крайне ответственная; она требует высокой лингвистической квалификации, лексикографического опыта, умения грамотно использовать технику количественного наблюдения. Среди составителей, консультантов и кураторов мы видим имена известных и авторитетных специалистов.

Важной особенностью рассматриваемых ЧС является их серийность. Получила окончательное, по-видимому, признание та точка зрения, согласно которой многоцелевой ЧС, особенно претендующий на охват большой жанрово-стилистической совокупности текстов, должен строиться как комбинация отдельных ЧС, имеющих самостоятельную применимость[1]. «Дифференциально-интегральный» подход в статистической лексикографии допускает типологическое сопоставление отдельных подъязыков, стилей, «регистров», он позволяет максимально разнообразить выборку, нивелировать неизбежные субъективные оценки в подборе текстов, нейтрализовать влияние поспешных решений в организации исследования.

ЧС рассматриваемых серий демонстрируют, далее, признание необходимости помещать в публикуемый словарь в с е лингвистические единицы данного уровня (слова и/или словоформы), которые обнаружены в обследованных текстах, в том числе и редкоупотребительные.

Всякий ЧС можно рассматривать, исходя из ряда свойственных любому из них признаков. К ним относятся прежде всего такие явные признаки, видимые при беглом взгляде на ЧС, как язык, входная единица, по-

[1] См., например, пятитомную серию ЧС латышского языка (1 — Техника и промышленность, 2 — Газеты и журналы, 3 — Художественная литература, 4 — Наука) [6—9], а также суммирующий материалы трех томов сводный том [10].

рядок размещения входных единиц, способ представления частот и других численных показателей. Дополнительную, не менее важную для характеристики словаря информацию сообщает составитель в комментарии или предисловии к ЧС, куда входят сведения о размере и содержании текстов и их распределении в выборочном корпусе, об объеме ЧС и, если не все его единицы вошли в публикацию, то о том, сколько их осталось за ее пределами, о технике анализа текстов, о принципах регистрации в тексте единиц ЧС и т. д.

С точки зрения полноты таких сведений серия ЧС белорусского языка выглядит наиболее экономной. Некоторые общие соображения о целях и смысле предпринятого исследования содержатся в предисловии к первому выпуску ЧС (художественная проза). Кроме него и вышедшего позднее (соответственно 1976 и 1979 гг.) ЧС публицистики запланирован ЧС деловых и научно-технических текстов. В дальнейшем предполагается анализ фольклорных и поэтических текстов, а также разговорной речи, в том числе и в том виде, в каком она представлена драмой и прозой.

Для первого ЧС использовались тексты послевоенных художественных произведений. Общее количество авторов и источников не указывается. Длина минимального фрагмента равна 1 тыс. словоупотреблений. Всего в корпусе текстов 290 таких фрагментов, а их суммарная длина составляет 290 словоупотреблений. В корпусе обнаружены 21 754 разных слова, за вычетом собственных имен и цифровых обозначений. Все эти слова помещены в ЧС в порядке убывания частот. Слова, встретившиеся не менее 15 раз, приводятся списком с указанием их частотных номеров, частот и количества фрагментов для каждого слова. Более редкие слова разбиты на группы по частоте и, внутри каждой группы, на подгруппы по количеству фрагментов текста.

1002 слова с частотами не менее 31 приведены дополнительным списком; при каждом из них указана его частота и число фрагментов, в которых оно встретилось.

Во введении не сообщается, как регистрировались словоупотребления текста, как они приводились к словарным формам. Поэтому читатель, знакомый с различными ЧС, может быть вначале озадачен, увидев именно слово *ён* («он») как самое частое (в художественной прозе, в письменной речи), а не такие слова, как *і* («и») или *у*, *ў* («в»). Недостающую информацию, правда, он может восполнить, просмотрев весь ЧС белорусского языка и обнаружив, что *ён*, очевидно, объединяет в себе все местоимения третьего лица обоих чисел. Однако повторять эту процедуру относительно подобных и других неясных случаев было бы затруднительно, поэтому составители должны были бы сами помочь в этом читателю.

Из-за того, что не указано соотношение прямой и авторской речи в выборке, можно предположить обязательным наличие первой в каждом фрагменте текста. Иначе трудно объяснить, почему местоимения *я* и *ты* встретились соответственно в 285 и 280 фрагментах из 290.

Второй выпуск ЧС (публицистика) составлен по выборке несколько большей. В 310 текстовых фрагментах по 1 тыс. словоупотреблений из ряда газет и одного журнала встретились 18 319 слов, не считая собственных имен. По оформлению оба выпуска идентичны. В частотном порядке представлены все слова, в алфавитно-частотном — около тысячи самых частых. Отличается второй выпуск от первого не только характером текстов, но и тем, что в перечень слов включены подвергнутые буквенной расшифровке числительные, которые в тексте употреблены в цифровой записи.

Хочется надеяться, что составители ЧС белорусского языка и их издатели не задержатся с публикацией очередных выпусков, а в этих выпусках желательно увидеть больше сведений о составе выборки и о способах ее анализа.

ЧС современного польского языка задуман и выполнен как серия из пяти отдельных словарей (планируется также выпуск сводного ЧС), изданных в виде пяти томов, четыре тома сброшюрованы в две книги каждый, а один — в три. Им предшествовал пробный том, ЧС публицисти

ки [2], повторенный затем с некоторыми модификациями как третий том серии. Объем выборки для каждого из пяти томов равен 100 тыс. словоупотреблений, и общая длина всего корпуса текстов доведена до 500 тыс. словоупотреблений.

ЧС польского языка примечателен не только тем, что он издан как серия отдельных, составленных по унифицированной методике ЧС, но и другими особенностями. Во-первых, он, как и рассмотренный выше ЧС, является «полным»; в нем приведены все обнаруженные в выборке слова (и, как увидим далее, словоформы). Во-вторых, и это самое главное его отличие от подавляющего большинства других ЧС, в нем приводятся как слова, так и словоформы, что сделано опять-таки «нетрадиционным» образом — каждое слово полного алфавитно-частотного списка сопровождается перечнем всех его форм, зарегистрированных в тексте. При словах и словоформах указаны их частоты и цифровые, кодовые обозначения частей речи, а при словоформах также и коды грамматических категорий.

Удачно выбранный способ представления двух ЧС (слов и словоформ) в одном словаре и наличие полных сведений о лексико-грамматических и грамматических значениях входных единиц значительно повышают практическую и теоретическую ценность ЧС этой серии и выгодно отличают их от других известных словарей.

В I томе (Научно-популярные тексты) содержатся 32,9 тыс. словоформ и 15,3 тыс. слов. Соответственно для II тома (Мелкие газетные информационные материалы) эти цифры равны 30,6 тыс. и 14,7 тыс., для III тома (Публицистика) — 28,4 тыс. и 12,1 тыс., для IV тома (Художественная проза) — 32,7 тыс и 16,4 тыс., для V тома (Художественная драма) — 23,2 тыс. и 11,2 тыс. Эти данные, кстати, могли бы несколько прояснить вопрос о количественной мере аналитизма языка как отношении числа разных слов к числу разных словоформ в одном и том же тексте. Предложенная двадцать лет назад [11], эта оценка, как ни казалась она информативной, вызвала возражения уже потому, что численное отношение слов к словоформам меняется с изменением длины текста [3]. Поэтому сравнивать языки с помощью таких оценок, если они получены на материале выборок несопоставимых объемов, было бы некорректным. Теперь к этому можно добавить, что для оценок аналитизма не следует использовать данные ЧС, не сопоставимых также и по содержанию выборок, на которых они базируются. Нетрудно увидеть, что «мера аналитизма» на материале I—V томов ЧС польского языка равна соответственно 0,47; 0,45; 0,43; 0,50; 0,50.

Кроме алфавитно-частотного списка всех слов и словоформ, в каждом томе приводятся полный обратный список слов с указанием их частот и лексико-грамматических кодов и частотный список слов с частотами не менее 9. Эта цифра принята в качестве пороговой и, по мнению составителей, означает, что при выборке в 100 тыс. словоупотреблений обеспечивает вполне допустимую ошибку в определении частот, величина которой равна 65%. Отметим, что очень часто разные составители разных ЧС достаточно произвольно задают величины коэффициентов, допустимых ошибок, подставляя их в общем-то похожие одна на другую формулы. Эти различия, впрочем, не должны смущать читателя, адресата ЧС, но только в том случае, если ему предлагается п о л н ы й словарь, а не урезанная его часть, которая осталась после удаления «недостоверной» зоны.

Обратный словарь — явление пока еще редкое в лексикографии [4], а обратный частотный словарь не менее редок и в статистической лексикографии. Поэтому отрадно отметить, что составители польского ЧС поместили обратные списки слов в каждый из пяти томов серии. Еще более, разумеется, повысили бы ценность словаря обратно-частотные списки с л о в о ф о р м, что сделало бы всю серию поистине уникальной.

---

[2] Рецензия на этот том опубликована в ФН, 1975, № 3.

[3] Это отмечено и составителями ЧС современной украинской художественной прозы во введении к словарю.

[4] Но все-таки, видимо, не слишком редкое, если способно уже вызвать тревогу: «Досужие лингвисты, армия которых в наше время растет с огромной быстротой, придумали даже словари обратные... Голова идет кругом!» [12].

Составители серии, сознавая, по-видимому, умеренность объема выборок и желая их максимально разнообразить, пришли к наиболее разумному в их случае решению, которое выдерживается на протяжении всей серии. Для каждого ЧС использованы очень небольшие фрагменты текста — по 50 словоупотреблений из каждого источника; всего таких фрагментов по 2 тыс. для одного ЧС. Этот важный признак — длину минимального отрезка текста — любого ЧС следует иметь в виду, когда данные ЧС используются при сопоставлении с данными других ЧС, особенно когда ЧС разных языков привлекаются для типологических наблюдений. Больше число разных источников, разных авторов, разных ситуаций дает и большее число разных единиц ЧС в отличие от такого ЧС, который, хотя и базируется на выборке сходного объема, но отражает меньшее число ситуаций, текстов, авторов. Об этом иногда забывают и в количественной лингвистике, и там, где используют ее наблюдения.

Введение к каждому тому серии содержит достаточную информацию об инструкции по разметке текста, по его подготовке к вводу в ЭВМ, о качественной и количественной структуре выборочного корпуса текстов и о способах его комплектования.

ЧС польского языка является значительным вкладом в славянскую и общую статистическую лексикографию.

Особое место в статистической лексикографии несомненно займет ЧС современного украинского языка, первый выпуск которого (в двух томах) охватывает лексику художественной прозы [5]. Характеризуется этот ЧС большим объемом выборки, равным 500 тыс. словоупотреблений для каждого из шести выпусков [5], богатством лингвистической и лингвостатистической информации, тщательно разработанной методикой анализа текста и такой формой представления материала, которая удобна для читателей с самыми различными интересами. Он, кроме того, является первым отечественным ЧС, в котором входная статья содержит и само слово, и все его словоформы, встретившиеся в выборке.

Лингвистические, лексикографические и статистические аспекты концептуального и «технологического» аппарата подробно изложены во введении. Это избавляет читателя от необходимости восстанавливать недостающие сведения по словарю, как это бывает во многих других случаях.

В выборочный корпус расписанных вручную текстов вошли 500 фрагментов по 1 тыс. словоупотреблений каждый из 25 книг 22 украинских писателей, опубликовавших свои произведения в 1945—1970 гг. Всего зарегистрированы 33 391 слово и 86 284 словоформы за вычетом собственных имен, рассматриваемых отдельно, нелитературных слов, записанных иноязычными алфавитами слов и числительных в цифровой передаче.

Единицы ЧС представлены в нескольких списках. Основной список, алфавитно-частотный, содержит в себе слова с частотами не менее 2 и их словоформы. При каждом слове и каждой словоформе указываются абсолютная и относительная частоты в речи персонажей и в авторской речи, общая абсолютная частота, число источников из 25, число фрагментов из 500, средняя частота, полученная делением абсолютной на постоянное число 500, и величина стандартной ошибки средней частоты.

Эти численные показатели заслуживают некоторого комментария. Единицы ЧС в прямой и авторской речи регистрировались раздельно потому, что составители считают принципиальными различия в функционировании лексики в обоих случаях. Составители же белорусского и польского ЧС не сочли необходимым обратить на это внимание. При комплектовании выборочного корпуса украинских прозаических текстов выдерживалось требование, чтобы прямая речь занимала не более 1/3 всего текста; в результате на нее пришлось 28,2% всех словоупотреблений выборки.

Приводимые в ЧС количественные характеристики важны и информативны, кроме одной — средней частоты. Составители сообщают, что она «содержит информацию для статистического сопоставления слов и слово-

[5] Планируемые шесть выпусков упоминаются в [13].

форм между собой в одной или нескольких выборках» [5, с. 14]. Однако каждое значение средней частоты — это соответствующая величина фактической, абсолютной частоты, деленная каждый раз на одно и то же число. Сопоставление слов или словоформ в пределах выборки (не говоря о нескольких, поскольку их еще нет, да и будут ли они включать в себя каждый раз по 500 фрагментов из 25 книг 22 авторов?) легче и естественнее по абсолютным частотам (в случае нескольких выборок картина также не изменится, поскольку они будут, очевидно, все одинаковых размеров) Гораздо важнее последняя характеристика — оценка ошибки средней частоты, которая может дать некоторое представление о большей или меньшей равномерности появления той или иной единицы в совокупности текстов. Однако говорить о равномерности распределения в строго статистическом смысле было бы, как кажется, преждевременным, если учитывать только эту характеристику, хотя и дополненную показателями распространенности — количествами источников. Величина оценки средней частоты дает информацию о равномерности/неравномерности появления только тех единиц ЧС, которые имеют одинаковые частоты. Более содержательной могла бы быть, скажем, методика А. Жюйана, использующая нормированный] коэффициент вариации и в общем-то хорошо известная зарубежным и отечественным лингвостатистикам [6]. При входных единицах алфавитно-частотного списка отмечены лексические омонимы, даны отсылки на фонетические варианты, начинающиеся с другой буквы, обозначения частей речи, кроме прилагательных, причастий и форм на *-но*, *-то*, и грамматических категорий в случаях грамматической омонимии.

Еще один список содержит словоформы одноразовых слов в алфавитном порядке. Они сопровождаются лексико-грамматической и грамматической характеристиками и указанием на номер источника и номер фрагмента. В трех последних списках приведены слова и словоформы, расположенные по убыванию частот и встретившиеся не менее 10 раз в прямой и авторской речи по отдельности и в общем корпусе — вместе.

Трудно было бы требовать от ЧС, некоего, пока еще гипотетического, стандарта относительно их оформления и представления числовой информации, особенно факультативной, и тем не менее ЧС современной украинской художественной прозы может рассматриваться сегодня как одна из образцовых работ в славянской и общей статистической лексикографии.

ЛИТЕРАТУРА

1. *Алексеев П. М.* К основам статистической лексикографии.— В кн.: Проблема слова и словосочетания: Межвузовский сб. научн. тр. ЛГПИ им. А. И. Герцена. Л., 1980.
2. *Алексеев П. М.* Статистическая лексикография (типология, составление и применение частотных словарей): Учебное пособие. Л., 1975.
3. *Мажэйка Н. С., Супрун А. Я.* Частотны слоўнік беларускай мовы. Мінск, 1976—1979.
4. Słownictwo współczesnego języka polskiego. Listy frekwencyjne. T. I—V. Warszawa, 1974—1977. (I — Teksty popularnonaukowe, 858 s.; II — Drobne wiadomości prasowe, 752 s.; III — Publicystyka, 684 s.; IV — Proza artystyczna, 885 s.; V — Dramat artystyczny, 631 s.).
5. Частотний словник сучасної української художньої прози. Т. 1—2. Київ, 1981.
6. *Jakubaite T., Krustovska D., Ozola V., Prāse R., Sika N.* Latviešu valodas biežuma vārdnīca. I sej. Tehnika un rūpniecība. 1. Daļa. Rīga, 1966.
7. *Jakubaite T., Guļevska D., Ozola V., Prāse R., Rubīna A., Sika N.* Latviešu valodas biežuma vardnīca. II sēj. Laikraksti un žurnāli. 1. Daļa. Rīga, 1969.
8. *Jakubaite T., Guļevska D., Ozola V., Rubīna A., Sika N.* Latviešu valodas biežuma vārdnica. 3 sēj. Daiļliteratūra 1. Daļa. Rīga, 1972.
9. *Jakubaite T., Grāvīte M., Ozola V., Rubīna A., Sika N.* Latviešu valodas biežuma vārdnīca. 4 sēj. Zinātne. Rīga, 1976.
10. *Jakubaite T., Ozola V., Rubīna A., Sika N.* Latviešu valodas biežuma vārdnīca. Apvienotais (1—3) sēj. Rīga, 1973.
11. *Пиотровский Р. Г., Алексеев П. М., Чернядьева Е. А.* Статистика речи и закономерности языка: Тез. докл. межвузовской конференции на тему «Язык и речь» (27 ноября — 1 декабря). М., 1962.
12. *Киселевский А.* «Комик загнул?» — Лит. газета, 1982, 26 мая.
13. Частотный словарь современной украинской художественной прозы (Пробная тетрадь). Киев, 1969, с. 1.
14. *Андрющенко В. М.* Новые работы в области статистической лексикографии.— ВЯ, 1968, № 5.

[6] Она подробно прокомментирована в [14].