

РОГОЖНИКОВА Р. П.

МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА И СЛОВАРНОЕ ДЕЛО

В соответствии с толкованием слова «фонд» в толковых словарях под Машинным фондом русского языка можно понимать ресурсы, запасы единиц русского языка, введенные в память ЭВМ и извлекаемые оттуда по мере необходимости. Машинный фонд можно рассматривать и как собственно фонд, хранилище разного рода данных, и как исследовательскую лабораторию, в которой на основе введенных в ЭВМ данных русский язык изучается в различных аспектах.

Наиболее актуальным, хотя и трудоемким, является создание машинного фонда лексических единиц. Актуально это потому, что именно на основе собрания лексических материалов возможна организация словарной работы, создание больших, фундаментальных словарей русского языка, а на их основе — кратких общих и учебных словарей. Именно словарная работа, особенно связанная с большими словарными предприятиями (такими, как словарь национального языка), требует наиболее полного охвата лексического состава русского языка, привлечения большого количества текстов произведений самых различных жанров, использования словарей как современного языка, так и предшествующих эпох, материалов русской разговорной речи, лексикологических исследований отдельных слов и т. п. С другой стороны, собрание лексических материалов русского языка дает возможность проводить разнообразные исследования лексики, ее семантических, грамматических, стилистических, словообразовательных и других особенностей. Поскольку слово — основная единица языка, собрание лексики может служить источником самых разнообразных работ, словарных и несловарных.

Создание машинного фонда лексических единиц возможно, с одной стороны, путем введения в ЭВМ текстов, обработки текстов таким образом, чтобы можно было выделить и подготовить для лексикографической обработки лексические и другие единицы текста. С другой стороны, машинный фонд предполагает обработку и хранение уже выделенных и лексикографически обработанных словарных единиц, т. е. таких, которые заключены в словарях русского языка, в лексикологических исследованиях.

Словарная работа обычно начинается с создания картотек. Создание картотеки — это сбор и подготовка языковых фактов к дальнейшей работе. Языковой факт представляет собой употребление слова в ограниченном контексте, извлечение слова из какого-либо текста, где выявляется его словарная форма, грамматические, стилистические особенности, словесное окружение, сочетаемость и т. п. В своей совокупности языковые факты представляют собой лексико-фразеологический состав русского языка.

Можно сказать, что у нас существует фонд русской лексики по источникам XIX—XX вв., правда, составленный вручную. Это Большая картотека Словарного сектора в Ленинграде. То же можно сказать и о картотеках исторических — Словаре русского языка XI—XVII вв., XI—XIV вв. в Москве и XVIII в. в Ленинграде, представляющих фонд русской лексики по указанным временным периодам. Каждая из этих картотек создавалась в ходе подготовки материалов для словаря той или иной эпохи. Организация картотек велась по определенному плану, связанному с замыслом словаря. В то же время каждая из картотек используется не только для непосредственной словарной работы, но и для многих других работ, словарных и несловарных. Широта охвата материала, большое количество разнообразных материалов на каждый языковой факт позволяет сделать

это. Так, на основе Большой картотеки Словарного сектора созданы не только Большие толковые словари (17-томный и 4-томный), но и Словари синонимов (2-томный и однотомный), словарь «Трудности словоупотребления и варианты норм русского литературного языка», завершен Фразеологический словарь русского языка. Материалы картотеки используются при подготовке Словарей новых слов, Словаря русских народных говоров. Здесь постоянно работают и другие исследователи, в том числе и иностранные ученые. Ведется работа по сбору материалов для кандидатских и докторских диссертаций, для монографических исследований, сбор и проверка материалов для учебных пособий, каталогов, учебных словарей, для плановых исследований Института русского языка, Института русского языка им. Пушкина, университетов страны и других научных и учебных учреждений. Картотекой пользуются литературоведы для комментирования текстов, изучения языка и стиля писателей и для других целей.

Создание машинного фонда лексики, на основе которого будут проводиться словарные, а также и несловарные работы, как представляется, целесообразно связывать с большим словарным предприятием — созданием словаря национального языка какого-либо временного периода. Об этом свидетельствует зарубежный и отечественный опыт [1—3], хотя имеются примеры создания центров по обработке больших массивов текстовых данных для различных лингвистических целей [4, 5].

Хотелось бы обратить внимание на те сложности, которые возникают при подготовке словарных картотек и которые следует учесть при создании Машинного фонда русского языка. В материалах картотеки необходимо по возможности максимально точно отразить лексическую систему русского языка, представленную через отдельные языковые факты. Сложность этой работы состоит в том, что она связана с охватом большого материала, огромного количества лексических единиц, нередко многозначных, которые нужно представить в возможно большей полноте, т. е. отразить лексический состав «вширь». Это можно сделать, привлекая большое количество разнообразных источников различных жанров. С другой стороны, каждый факт языкового употребления должен быть обеспечен достаточным количеством материалов. Это необходимо для того, чтобы выводы, которые будут делаться в отношении семантики отдельного слова или группы слов, сочетаемости, грамматических и других особенностей слова, не были случайными. Таким образом, необходимо отразить состояние лексико-фразеологической системы «вглубь». Это может быть достигнуто при тщательной, возможно более полной обработке источников.

И полнота отражения лексических единиц, и обеспеченность их достаточным количеством материала позволят более объективно включать различные группы лексики в словники словарей и отразить каждое языковое явление в словаре с наибольшей точностью. Исходным материалом являются письменные тексты, которые отбираются в зависимости от целей и задач словаря [6]. Они представляют русский литературный язык во всем его жанровом и стилистическом многообразии. Отбор текстов для картотеки — серьезный и важный вопрос, которому уделяется большое внимание.

Что касается источников картотеки словаря современного языка, то они должны быть значимыми и актуальными, в них должен быть представлен русский литературный язык во всех важнейших сферах его функционирования. Необходимо, чтобы они были достаточными в количественном отношении, поскольку важно выявить реальный лексический состав русского литературного языка и обосновать языковой факт необходимым количеством материала (контекстов). При выборе источников учитывается принадлежность к определенному стилю и жанру, тематика и языковые качества источника. Только большое количество разнообразных источников может отразить состояние литературного языка в данный временной период.

В настоящее время в словарной картотеке, на основе которой подготовлены такие фундаментальные словари, как 17-томный и 4-томный, а также составлен и составляется ряд других словарей, насчитывается более

7 млн. карточек-цитат. Это крупнейшее собрание лексических материалов. Казалось бы, можно перевести на магнитные ленты картотеки, созданные вручную, и это создало бы машинный фонд лексики русского языка. Однако в картотеках, составленных вручную, имеются, как правило, уязвимые места, связанные с тем, что пополнение картотек происходит неравномерно в разное время, основным принципом выборки является частичная выборка, ориентированная прежде всего на качество цитаты, т. е. на непосредственное использование ее в словарях. Не учитывается весь реальный лексический состав хотя бы какого-то специально отобранного круга источников. Все эти недостатки картотек объяснимы. Они связаны с тем, что обработка источников — очень трудоемкое дело — происходило, да и сейчас еще происходит вручную. Это не дает возможности учесть весь реальный лексический состав источников, не всегда имеется достаточное количество материалов на отдельный языковой факт.

Использование ЭВМ для подготовки словарных материалов поднимет эту работу на более высокий уровень. На первых порах кажется целесообразным сочетание ручных и автоматизированных методов создания картотек (т. е. пополнение картотек, созданных вручную, данными, полученными на ЭВМ) и одновременно с этим разработка программ и подготовка текстов для машинного фонда лексики. Создание Машинного фонда русского языка связано с использованием современной сложной техники, знанием специальных языков, с помощью которых производится составление программ, позволяющих получить необходимые результаты на естественном языке, и т. д. Все это предполагает разработку специальных задач, относящихся к разным сторонам лингвистики. Эти задачи решаются специалистами по вычислительной технике совместно с лингвистами [7].

В настоящее время значительную долю времени, связанного с обработкой текстов на ЭВМ, занимает подготовка текстов для ввода. Для повышения эффективности обработки текстов на ЭВМ целесообразно использование уже готовых перфолент, магнитных лент и т. п. Это экономит время подготовки информации для ЭВМ и позволяет ввести в ЭВМ текст, адекватный тексту источника. Текст, введенный в память ЭВМ, может быть отредактирован.

Возможности редактирования машинного текста в режиме непосредственного доступа дают лексикографу дополнительные удобства при работе с источником. Он может оперативно менять издание, выбирая первое или наиболее авторитетное или любое другое в зависимости от целей обработки текста. Например, если обрабатывается текст произведений XIX в., изданный в XX в., то введя его один раз, можно работать с ним дважды — по изданию XIX и XX вв.

Языковые явления и факты поступают из ЭВМ в виде алфавитно-частотных указателей словоформ — на листах, в виде конкордансов, алфавитно-частотных указателей на карточках и т. п. Получение словоуказателей текстов, конкордансов проводится и у нас, и за рубежом [8, 9]. Много сделано в этом отношении теми, кто занимался подготовкой частотных словарей и вопросами машинного перевода [10, 11]. В алфавитно-частотном указателе словоформ приведены все словоформы текста с их вариантами, морфологическими и другими особенностями, даются количественные указания использования их в тексте. Все это очень важно для лексикологов, лексикографов. На основе прямого указателя словоформ составляется словоуказатель текста, в котором все словоформы приведены к основной форме слова. Он необходим для сопоставления данного текста со словарями, картотеками и т. п., где за основу берется исходная форма слова.

Автоматическое приведение словоформ к основной (словарной) форме слова, или лемматизация, затруднительно из-за большого количества фоно-морфологических особенностей лексики, представляющих национальные отличительные черты русского языка. Ср., например, чередование в корне слова, аффиксах (*носить* — *ношу*, *разбить* — *разобью*). Кроме того, приведение к основной форме слова выявляет большое количество омоформ и омографов, которые могут быть заданы только списком (напри-

мер: род. п. ед. ч. существительного: *бег* — *бегу* и форма 1-го л. ед. ч. глагола: *бежать* — *бегу*). Поэтому после составления указателя словоформ каждой словоформе в интерактивном режиме приписывается его словарная форма, а затем производится пересортировка в алфавитный порядок по словам в словарной форме [12]. Для автоматизации этого процесса необходимо создание списка соответствий, в который войдут все словоформы, соотносящиеся с основной словарной формой слова и не имеющие каких-либо особенностей (*делаю* — *делаю*), омоформы и омографы (*бегу* < *бег* и *бежать*) и формы с чередованием в корне или аффиксах (*разобью* — *разбить*). Использование такого списка облегчит лемматизацию и сортировку слов при автоматической обработке текста.

Возможно также приписывание к каждой словоформе основной формы слова в исходном тексте, но она целесообразна лишь на небольших по объему текстах. Задача составления списка омоформ и омографов представляется важной не только в узко практическом (для машинной обработки текстов), но и в теоретическом отношении, а также для обучения русскому языку [13—15]. В настоящее время работа по выявлению и составлению списка омографов ведется в Картотеке Словарного сектора.

Словоформы, приведенные к основной форме слова, могут быть выданы как списком, так и на карточке вместе с количественными данными употребления словоформ и всего слова в тексте. Возникает возможность организовать картотеку-словоуказатель, являющуюся необходимой частью словарной картотеки, но, как правило, отсутствующую в современных картотеках, поскольку создание ее является очень трудоемким делом. Это по существу словник с приведением всех словоформ к основной форме слова и количественными данными их употребления в полностью обработанных текстах. Такая картотека-словоуказатель представляется весьма важной с разных точек зрения: она дает представление о словнике по полностью обработанным произведениям. Здесь воспроизведен весь набор словоформ текстов, даются сведения о количественных показателях употребления словоформ и слов. Картотека-словоуказатель позволяет пополнять картотеку цитат в необходимых случаях по частям, дает возможность быстро найти нужную цитату в тексте, а также использовать количественные данные об употреблении слов и словоформ в текстах. Гораздо легче проводить сверку словника со словарями по картотеке-словоуказателю, поскольку здесь все словоформы, как и в словаре, приведены к основной форме слова.

Получение карточек-цитат из произведений, введенных в ЭВМ, — следующий этап автоматизации. Контекст выделяется формальным путем: по количеству словоупотреблений, строк от заданного слова. В стихотворных произведениях — это строфа, которая обычно представляет законченный отрезок в смысловом отношении, для прозаических произведений — определенное число символов вправо и влево от ключевого слова, с учетом пунктуационных знаков [16]. На карточке имеются библиографические сведения. Таким образом, лексикограф-исследователь получает по своему требованию любые цитатные материалы тех произведений, которые обработаны с помощью ЭВМ. Обработка текстов с помощью ЭВМ предполагает создание программы для получения обратных указателей словоформ и обратных словоуказателей. Они могут быть использованы для выявления вариантных форм и слов и сбора материалов для них. Обратный указатель словоформ дает полную картину словоизменения, которую невозможно найти в обычном обратном словаре, в основу которого положены исходные формы слова. В то же время обратный словоуказатель, составленный по исходным формам слов, может послужить основой для выделения и детального обследования определенных групп лексики, исходя из их словообразовательных признаков.

Большое значение имеет полная и последовательная фиксация устойчивых и фразеологических сочетаний в словарях. Эта задача является сложной как с лингвистической, так и с технической стороны. Необходимо знание законов построения сочетаний, их формальной структуры. Формальным путем выделяются сочетания с постоянным порядком слов. Выделение

их производится по опорным словоформам [17], а также по определяемым формальным путем левым и правым границам. Работа по выявлению сочетаний с постоянным порядком слов способствовала появлению учебного «Словаря сочетаний, эквивалентных слову» [18]. Интересной представляется методика выделения именных словосочетаний, предложенная для технических текстов [19].

Обработка текстов с помощью ЭВМ даст возможность более полно отразить в толковых словарях лексический состав русского языка, что окажет влияние и на другие стороны исследований по русскому языку. Так, в толковые словари по разным соображениям, в том числе и из-за недостаточного количества материалов, не включаются целые группы лексики (например, наречия на *-о*, *-е*, наречия с *по-* и др., слова с аффиксами субъективной оценки и т. п.). Это характерно не только для русской лексикографии.

Работа по созданию машинного фонда лексики безусловно является трудоемкой, но она открывает новые возможности для словарной работы, поскольку обрабатываются полностью произведения определенного временного периода. Для этой цели необходимо обрабатывать такие материалы, которые могут быть использованы многократно для разных работ. Так, если взять произведения классической литературы XIX в., не только художественной, то они могут быть использованы в словарях современного русского языка, общих и учебных, для создания словаря языка XIX в., для различного рода лингвистических и других исследований. Еще Я. К. Грот и А. А. Шахматов при работе над словарем русского языка высказывали идеи о необходимости создания словарей языка наиболее крупных писателей, что дало бы возможность гораздо полнее отразить лексический состав русского языка в словаре [20, 21].

Другая важная часть словарной работы — отбор лексики в толковые словари. Сверка словника картотеки со словниками словарей — важный этап работы при формировании словников словарей. В настоящее время в распоряжении лексикографов имеются различные словари, позволяющие обобщенно характеризовать объект исследования — слово. Однако пока еще нет сводного справочника, в котором можно было бы найти ответ на вопрос: в каких существующих словарях зафиксировано то или иное слово. Поскольку словники различных, в том числе и толковых словарей русского языка не совпадают, представляется целесообразным создание справочника, в котором бы все словники наиболее значительных словарей были объединены. Таким справочником должен явиться Сводный словник словарей русского языка. Он будет объективным справочником, с помощью которого можно формировать словники различных по объему словарей, как общих, так и учебных. Разнообразие задач, которые могут решаться с помощью Сводного словника словарей, обуславливает его многократное использование в различных целях. Сводный словник словарей русского языка может стать основой для Генерального словника в машинном фонде лексики, создаваемом с помощью словарей.

На основе имеющихся полностью обработанных текстов и частотных словарей произведений авторов XIX в. предполагается создание Словаря редких слов. Сопоставление этого словаря со Сводным словником словарей выявляет слова, не включенные в современные толковые и другие наиболее значительные словари русского языка, но употреблявшиеся в произведениях русской классической литературы [22]. Этот словарь может быть составлен на основе полностью обработанных на ЭВМ текстов произведений авторов XIX в. В настоящее время по имеющимся словоуказателям и полным частотным словарям авторов XIX в. выявлено около 3000 таких слов. Эта работа важна для лексикографов, литературоведов. Она помогает увидеть жизнь слов.

Сводный словник словарей имеет и самостоятельное научное значение. Он помогает проанализировать состав словников словарей, на основе которых он создан. По Сводному словнику можно определить, какой лексикой пополнялись словари, будут пополняться в дальнейшем, какие слова исключались из словарей.

Известно, что в настоящее время терминологическая лексика получает значительно большее распространение в общелитературном языке. В филологических словарях терминологическая лексика могла бы быть представлена значительно полнее, если бы в большей степени учитывались не только данные картотеки, но и данные различных словарей. Данные терминологических словарей и словарей новых слов позволят с большей достоверностью вводить их в толковые словари. В то же время Сводный словник показывает необходимость нормализации терминологической лексики в нефилологических словарях. Нередко одно и то же слово в разных нефилологических словарях представлено по-разному. В связи с этим целесообразно создание нормативного Словаря межотраслевой русской терминологии.

С помощью Сводного словника можно выявить наиболее употребительную лексику современного русского языка и отобрать ее в словари различных объемов. Это важно при составлении толковых словарей разного объема как общего, так и учебного типа, а также для отбора лексики в русско-иноязычные словари. В настоящее время у нас отсутствуют толковые словари на 10, 20, 30 тыс. слов, необходимые как для русских, так и для тех, кто изучает русский язык на продвинутых этапах. Не очень ясны пока критерии отбора лексики в такие словари. Думается, что Сводный словник мог бы оказаться полезным и в этом отношении. Поскольку в Сводном словнике отражены варианты слов, отмеченные в словарях, а также омонимы, некоторые грамматические особенности слов, он может оказать существенную помощь в научных исследованиях и учебных занятиях по лексикологии и служить базой для отбора лексики по различным семантическим, грамматическим и другим группам, для изучения омонимов, вариантов слов. По типу Сводного словника словарей современного языка могут быть созданы словники словарей предшествующих эпох, терминологических словарей и т. п.

С созданием словарей связана еще одна сторона работы, которая также должна войти в Машинный фонд русского языка. Речь идет об учете лексикологических работ, в которых в том или ином аспекте исследуются лексические единицы.

При каждом слове дается библиографическая справка, в которой указывается кем, когда и где опубликована работа о данном слове, с краткой аннотацией. Подобный указатель по истории слов русского литературного языка XVIII—XX вв. уже имеется [23]. Сейчас в Словарном секторе ведется многоаспектное обследование лексики, зафиксированной в существующих лексикологических трудах.

Введенные в ЭВМ сведения позволяют поставить вопрос о включении в словари определенных разрядов лексики, об изменении толкований слов, дадут возможность уточнить или изменить стилистическую квалификацию слова, этимологическую справку и т. п. Возможность использования ЭВМ в лингвистических исследованиях с каждым годом расширяется в связи с усложнением функций современных вычислительных машин, увеличением объема памяти и быстродействия, оснащенностью дополнительными устройствами, облегчающими ввод и обработку исходного материала (текстов или словарей). Применение интерактивных методов позволяет исследователю-лингвисту использовать ЭВМ даже в тех случаях, когда какие-то языковые факты или явления трудно или невозможно формализовать. Желательно передать ЭВМ наиболее трудоемкие и часто повторяющиеся задачи, что позволит значительно сократить сроки исполнения различных разработок, связанных с исследованием языка, и в то же время откроет новые возможности его изучения.

ЛИТЕРАТУРА

1. Josselson H. H. Lexicography and the computer. — In: To honor Roman Jakobson. V. II. The Hague — Paris, 1967.
2. Robinson J. L. Bailey R. W. Computer-produced microfilm in lexicography: toward a dictionary of Early modern English. — In: The computer and literary studies. Ed. by Aitken A. J. Edinburgh, 1973.

3. Virks J. An automatic system for generation of dictionaries. In: Computational linguistics and related topics. Tallin, 1980.
4. Meunier J. G. A system for text and content analysis.— Computers and the humanities, 1976, v. 10, № 5.
5. Hage J. H. A survey of EDP projects in linguistics and literary studies at Norwegian universities.— ALLC bulletin, 1975, v. 3, № 3.
6. Левашов Е. А., Рогожникова Р. П. Принципы отбора источников для картотеки Словаря современного русского литературного языка.— В кн.: Теория и практика современной лексикографии. Л., 1984.
7. Вычислительная лингвистика. М., 1980, 1982.
8. Рогожникова Р. П., Чернышева Л. В. Организация словарной картотеки на базе автоматизированной системы.— В кн.: Теория и практика современной лексикографии. Л., 1984.
9. Герд А. С., Богданов В. В., Азарова И. В., Аверина С. А., Зубова Л. В. Автоматизация в лексикографии и словари-конкордансы.— ФН, 1981, № 1.
10. Статистика речи и автоматический анализ текста. Л., 1968, 1972, 1974.
11. Автоматическая переработка текста методами прикладной лингвистики. Кишинев, 1977.
12. Koubourlis D. J. From a word-form concordance to a dictionary-form concordance.— In: Computers in the humanities. Minneapolis, 1974.
13. Виноградов В. В. Об омонимии и смежных явлениях.— В кн.: Виноградов В. В. Избранные труды. Исследования по русской грамматике. М., 1975.
14. Мельникова А. И. К вопросу о русских омографах.— РЯШ, 1974, № 4, с. 60.
15. Колесников Н. П. Изучение проблем омонимии в советском языкознании.— В кн.: Вопросы русского языкознания. Куйбышев, 1978.
16. Tollenaere F. The problem of the context in computer-aided lexicography.— In: The computer and literary studies. Ed. by Aitken A. J. et al. Edinburgh, 1973.
17. Данейко М. В., Машкина Л. Е., Нехай О. А., Сорокина В. А., Шаранда А. Н. Статистическое исследование лексической дистрибуции словоформ.— В кн.: Статистика речи. Л., 1968.
18. Рогожникова Р. П. Словарь сочетаний, эквивалентных слову. М., 1983.
19. Белоногов Г. Г., Загика Е. А., Новоселов А. П., Хорошилов А. А. Автоматическая нормализация слов и словосочетаний.— НТИ, сер. 2, 1985, № 1, с. 39—40.
20. Грот Я. К. Филологические разыскания. Т. I. Материалы для словаря, грамматики и истории русского языка. СПб., 1876, с. 160—172.
21. Истрина Е. С. А. А. Шахматов как редактор Словаря русского языка.— ИАН ОЛЯ, 1946, № 5, с. 405—407.
22. Рогожникова Р. П. Редкие слова в произведениях авторов XIX в.— ВЯ, 1982, № 1.
23. История слов русского литературного языка (XVIII—XX вв.): Опыт указателя литературы, изданной в СССР на русском языке с 1918 по 1970 г. Сост. Левашов Е. А., ред. Сорокин Ю. С.