ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 6 1986

ГЕРД А. С.

РУССКАЯ МОРФОЛОГИЯ И МАШИННЫЙ ФОНД РУССКОГО ЯЗЫКА

Машинный фонд русского языка (далее — МФРЯ) — качественно новая филологическая проблема, поставленная перед современным теоретическим и прикладным языкознанием. Разработка, окончательное создание и постепенный ввод в эксплуатацию МФРЯ в единой цепи комплексной автоматизации научно-исследовательских работ в нашей стране, включая лингвистические институты и кафедры, во многом окажется поворотным моментом в дальнейшем развитии науки о русском языке и использовании ее результатов в общественной, культурной и народно-хозяйственной деятельности.

До последнего времени большинство научных разработок в области прикладной и математической лингвистики было ориентировано на различные чисто практические направления, связанные с народным хозяйством [5]. МФРЯ позволит направить достижения современной прикладной лингвистики в области автоматической переработки текста и на решение собственно филологических проблем.

Создание МФРЯ — фундаментальная гуманитарная задача, которая потребует в перспективе и большой нравственно-психологической перестройки кадров филологов. Именно поэтому нам представляется целесообразным проведение широкого обсуждения в печати общефилологических, чисто лингвистических, технологических, юридических, правовых, психологических и организационных аспектов создания МФРЯ. Только в результате такого обсуждения и последующего анализа различных мнений можно будет выработать правильную стратегию разработки МФРЯ в целом.

Согласно публикациям [1, 4], МФРЯ должен содержать: «генеральный словник русского языка; иллюстрационно-текстовый фонд русского языка; терминологический фонд русского языка; академический словарно-грамматический фонд русского языка; лексикографическую базу Машинного фонда русского языка; лингвостатистическую базу Машинного фонда русского языка; фонд процессоров русского языка; фонд лингвистических алгоритмов и программ; информационно-справочный фонд по русистике» [4, с. 55]. Актуальность создания в МФРЯ таких фондов, как генеральный словник, терминологический фонд, иллюстрационно-текстовый фонд, в принципе не вызывает сомнений, хотя каждый из них ждет своего особого рассмотрения.

Более серьезного обсуждения требует словарно-грамматический фонд, при этом в комментариях нуждаются обе части определения «словарно-грамматический». Однако если о словарной части этого фонда в названных статьях говорится довольно подробно, то о его грамматической части сказано более лаконично: «Представляется целесообразным непосредственно связать собственно словарные базы с базой данных Академической грамматики, поместив в качестве адресующего индекса к текстам Грамматики ее словарный и предметный указатели, через которые нужные места Грамматики свяжутся с нужными местами словарных статей» [1, с. 7].

В статье В. М. Андрющенко [4] высказывается мысль о необходимости создания справочного фонда русских академических грамматик. Здесь далеко не все ясно, многое в этих положениях нуждается в раскрытии и уточнении. На наш взгляд, именно обсуждение вопроса о представлении

морфологии русского языка в МФРЯ позволит отчетливее представить специфику и место других подфондов такого типа в МФРЯ в целом.

Между такими фондами, как генеральный словник или иллюстрационно-текстовый фонд, и подфондом грамматики (грамматик) русского языка есть существенная разница. Первые типы фондов — это собрание самих фактов языка (слов, предложений), которые на современном уровне автоматической переработки относительно нетрудно извлечь непосредственно из самих текстов. Фонды грамматики — массивы совсем другого рода, это фонды интерпретационные, которые всегда предстают перед нами уже как итог научно-исследовательской деятельности языковедов. При этом морфология русского языка существует в наши дни в двух основных видах. Во-первых, в виде грамматической теории, которая развивалась и развивается вне прямой и непосредственной ориентации на автоматическую переработку текста, и, во-вторых, в виде машинной и предмашинной морфологии, возникшей с конца 50-х годов XX в. Учитывая это, рассмотрим перспективы организации подфонда русской морфологии в МФРЯ. В этой связи необходимо поставить прежде всего вопрос: «Кому, для чего и какого рода справки, ответы сможет выдавать морфологический подфонд МФРЯ?». В самом общем виде на этот вопрос можно ответить так: «Каждому, кому в тех или иных теоретических или практических целях понадобятся сведения по морфологии русского языка».

Не касаясь всех проблем морфологии современного русского языка, рассмотрим поставленный вопрос на материале только таких ее аспектов, как классификация частей речи, морфемика, словообразование и формообразование, и при этом сначала в том его виде, как он представлен в неалгоритмической, немашинной морфологии [1]. Предположим, что в систему МФРЯ поступил запрос: «Выдать список частей речи современного русского литературного языка». Такой запрос вполне может исходить от специалистов по русскому языку как иностранному, от методистов или работников издательств, отраслевых НИИ, разрабатывающих словари, справочники, классификаторы. Во-первых, для того чтобы потребитель получил ответ на этот запрос, в ЭВМ заранее формально должны быть заложены хронологические границы современного русского языка в области морфологии. В то же время очевидно, что эти границы в морфологии вряд ли будут полностью совпадать с аналогичными границами в лексике. Разными они могут оказаться и для различных функциональных жанров, типов текстов. Во-вторых, ответ на этот запрос предполагает, что в ЭВМ заранее уже введены параметры русского языка как литературного, которые и приписаны соответственно грамматическим категориям и формам. Из последнего, в свою очередь, вытекает, что в МФРЯ должны быть эксплицитно выделены категории и формы, не характерные, не типичные для русского литературного языка, и здесь вновь встанут проблемы о месте языка науки, о языке фольклора, о литературной разговорной речи и т. д. Нетрудно заметить, что ответ даже только на эти два аспекта одного запроса потребовал бы глубокого пересмотра целого ряда кардинальных проблем русской филологии.

Обратимся к тому, справку о каких частях речи выдаст потребителю МФРЯ. Казалось бы, что здесь ответить уже гораздо легче. Однако ответ ЭВМ на эту часть запроса прямо зависит от того, какая из концепций частей речи в нее введена.

Отнюдь не возвращаясь вновь к рассмотрению всех точек зрения по вопросу о частях речи в русском языке, отметим только, что между разными концепциями (Ф. Ф. Фортунатова, А. М. Пешковского, М. Н. Петерсона, А. А. Шахматова, Л. В. Щербы, В. В. Виноградова, Л. А. Булаховского или последними идеями порождающих грамматик) никакое промежуточное решение, применимое специально для МФРЯ, найдено быть не может [2].

---

[1] В статье не затрагиваются вопросы грамматической семантики.

[2] Общий обзор проблемы частей речи в русском языке см. в [6] (там же и обширная литература вопроса); см. также [7—9].

Если МФРЯ строится на научной основе, то он вынужден будет выдавать потребителю не абстрактный список частей речи вне времени и пространства, а живую концепцию, занимающую свое место в истории науки о русском языке. Нетрудно себе представить также, например, все те трудности, которые встанут перед разработчиками МФРЯ при выделении и формализации признаков служебных частей речи. Немало сложностей возникает и с чисто информационной стороны. Оказывается, что даже если в МФРЯ будут введены различные частеречные концепции, то потребитель без специального информатора, осведомленного о существующих школах и направлениях, либо вообще не сможет получить ответ на свой запрос, либо получит его в таком суммарном виде, в котором он сам не сможет разобраться, а специалиста-лингвиста вряд ли удовлетворит ответ, столь неполный с энциклопедической точки зрения.

Рассмотрим теперь типы возможных запросов в МФРЯ по морфемике и словообразованию. Это могут быть, например, такие запросы, как: 1) выдать полный список морфем русского языка; 2) выдать все словообразующие морфемы; 3) выдать все формообразующие морфемы; 3) выдать списки суффиксов (приставок); 5) выдать структурные типы слов; 6) выдать типы производных слов (суммарно или отдельно по каждому типу); 7) выдать списки слов такой-то лексико-семантической группы, образованных по тому или иному способу словообразования.

Однако и здесь даже ответ на казалось бы самый простой запрос типа «выдать список суффиксов русского языка» полностью зависит от концепций, заложенных в МФРЯ, и, в частности, от интерпретации в МФРЯ таких проблем, как значение морфемы, принципы морфологической сегментации текста, повторяемость и регулярность морфем, унификация, наличие или отсутствие у морфемы материальной оболочки, подвижность морфем, интерфиксация, линейно и нелинейно выделимые морфемы, фонология морфемы и т. д. По всем этим проблемам мы имеем в литературе по русистике порой совершенно различные подходы (Ф. Ф. Фортунатов, Д. Н. Ушаков, Г. О. Винокур, А. И. Смирницкий, Н. М. Шанский, Е. А. Земская и др.)[3].

Новые подходы в морфологии словообразования влекут за собой новое понимание способов словообразования, их места в общем процессе порождения нового слова. Более того, именно в словообразовании имеем немало случаев, которые нередко вполне допускают двоякое решение вопроса о том, как образовано то или иное слово (проблема множественности мотиваций). Это, например, многие слова с суф. *-ник-*, *-ница*, *-овик*, *-овка*. Ср.: *школьник*,— от *школа* или *школьный*, *колхозник* — *от колхоз* или *колхозный*, *грузовик* — от *груз* или *грузовой*, *купальник* — от *купаться* или *купальный костюм* и т. д.

Наконец, нередко встречается мнение, согласно которому в МФРЯ должны быть заложены различные словари морфем русского языка. Однако из сказанного выше очевидно, что любой словарь морфем есть отражение теоретической концепции его автора, и в зависимости от того, какая из этих концепций найдет место в МФРЯ, и будут выданы различные списки морфем и алломорфов. Выдавать же потребителю реестры морфем длиной в несколько сот единиц, в которых он потом должен сам как-то разобраться, вряд ли целесообразно.

Запросы по формообразованию типа «от какой основы, при помощи какой морфемы образована та или иная конкретная форма?» внешне могут показаться даже наивными по своей простоте. Однако каждый русист знает, сколько здесь неясного и порой действительно трудного именно на уровне анализа отдельных слов и форм.

Как показывает опыт разработки различных информационных систем, потребителю, специалисту обычно нужны сведения не общего характера, а справки о частном, о деталях, о свойствах и качествах объектов. Запросы типа «от каких основ образуются в русском языке полные причастия прошедшего времени?» будут гораздо более редки, чем запросы типа «как об-

[3] Общие обзоры проблем морфемики см. в [10—16].

разована форма инфинитива *жать* или форма 3-го л. мн. числа *гнут?*». Эффективность работы любой информационной системы определяется числом обращений к ней, которое, в свою очередь, будет расти или уменьшаться в зависимости от степени удовлетворенности потребителя результатами работы этой системы, от ее пертинентности. Вряд ли стоит говорить о том, что по мере обращения к таким разделам морфологии русского языка, как формообразование глагола, вид глагола или категория состояния, число подобных спорных вопросов будет возрастать в геометрической прогрессии.

Рассмотрим теперь, сможет ли МФРЯ отразить машинную морфологию. Машинная морфология русского языка — сегодня довольно солидная совокупность различных словарей основ, суффиксов, устойчивых оборотов и алгоритмов их обработки.

Широкое распространение в свое время нашел строго формальный алгоритмический подход выделения частей речи, парадигм и морфем, разработанный Н. Д. Андреевым и его последователями [17—20]. Р. Г. Piотровским и его учениками были созданы по единой методике различные алгоритмы морфологического анализа разработаны принципы выделения частей речи, основ, флексий, все алгоритмы и программы отлажены и апробированы на ЭВМ в различных конкретных системах автоматической обработки текста. Создан сводный многоотраслевой автоматический словарь русского языка (МАРС), содержащий немало грамматической информации [21—23].

В ряде работ Б. Ю. Городецкого обоснована идея создания проблемно ориентированных инвентарей морфем языка (см., например [24]). Значительное число разработок в области собственно машинной морфологии принадлежит Г. Г. Белоногову и его коллегам: это — строгие, эффективно действующие автоматические словари словоформ, концов слов, окончаний, суффиксов, сочетаний суффиксов, словообразовательных классов слов, нормализации словоформ [25—28]. Широко используются сведения по морфологии в системах машинного перевода [29—32]. Обстоятельный общий обзор принципов автоматического морфологического анализа дан в [33].

Могут ли быть отражены и каким образом все эти сведения в МФРЯ? Как ни парадоксально, но и при обращении к чисто формальной машинной морфологии мы сталкиваемся с теми же проблемами, о которых уже говорилось. В различных алгоритмах морфологического анализа по-разному выделяются части речи, основы и аффиксы в одних и тех же типах.

Во многих исследованиях по машинной морфологии выделение частей речи строится только на основе частотных словарей, и если такой подход еще применим к собственно знаменательным частям речи, то он вряд ли оправдан при определении частеречной принадлежности слов типа, *едва*, *лишь*, *какой*, *как*, *где*, *тут* и т. п.

Применительно к морфемике и словообразованию создано немало алгоритмов вычленения основ и аффиксов, но и здесь при использовании разных алгоритмов выделяются различные основы и аффиксы, по-разному проводятся разграничительные линии между словообразованием и словоизменением [4].

К сожалению, в отдельных автоматических словарях принципы формализации морфологии не оговорены. Наконец, логично было бы предположить, что реестры основ и аффиксов, разработанные в машинной морфологии, могут понадобиться во многих практических приложениях, связанных с автоматической переработкой текста. Однако специалисты по прикладной лингвистике знают, что практически сегодня каждая конкретная автоматизированная информационная система использует свой алгоритм морфологического анализа, а многие информационные системы работают вообще без привлечения морфологии, «минуя морфологию». Большинство же универсальных алгоритмов, как известно, быстро дает сбой

[4] Аналогично весьма трудно представить себе непротиворечивую и свободную от исследовательских возарений информационную систему по этимологии.

при переходе на большие объемы текстов или к новым их типам. И в этом отношении ждать от МФРЯ неких универсальных алгоритмов морфологической обработки текста нереально, имея в виду как теорию, так и перспективу появления новых поколений ЭВМ. Новые задачи, новые типы машин потребуют и новых конкретных подходов к морфологии русского языка.

Таким образом, и здесь МФРЯ не сможет выдать никаких единых и непротиворечивых реестров основ, суффиксов, алгоритмов морфологического анализа, независимого от авторской концепции.

Естественно возникает вопрос: «Возможно ли вообще создание морфологического подфонда МФРЯ?» И возможно, и нужно. Но морфологический подфонд МФРЯ должен разрабатываться на принципиально иных основах, чем такие фонды, как генеральный словник или иллюстрационно-текстовый фонд.

Во-первых, в рамках МФРЯ должна быть создана специальная документально-фактографическая система по морфологии русского языка. Такая информационная система, не выдавая неких идеальных и универсальных списков морфем и способов слово- и формообразования, сможет дать потребителю гораздо больше.

Обращаясь к этой системе, мы узнаем, где, когда, кем и в какой работе (с ее точными выходными данными) рассматриваются, например, принципы выделения частей речи или морфем, формы русских глаголов или инфинитивы. Специалисты по методике русского языка, учителя смогут легко и быстро получить сводные аннотированные справки о новых работах по преподаванию русского языка в школе. В такой фонд обязательно должно быть введены многообразные внутриведомственные методические разработки и сборники упражнений, которые представляют поистине золотой фонд методики преподавания русского языка, но практически нигде никак не учитываются.

Разработка такой документально-фактографической системы по морфологии потребует привлечения квалифицированных специалистов по грамматике русского языка, ибо в конечном счете речь пойдет о выработке целостной шкалы морфологических параметров, на основе которых должна будет работать и сама система. Так, например, только для морфемики следует учесть такие признаки-параметры, как наличие в той или иной работе определения и описания а) морфемы; б) алломорфа; в) варианта морфемы; г) принципов выделения и идентификации морфем; д) дистрибуции морфемы; е) структурных типов слов и т. д. Каждая часть речи, каждая грамматическая категория потребуют детального выявления своих параметров, релевантных для потребителей разных типов. Списки таких параметров должны быть открытыми и доступными для пополнения.

Само собой очевидно, что разработка источниковедческих и теоретических основ такой системы — проблема чисто лингвистическая. Реестры параметров, признаков должны создаваться независимо от математического программного обеспечения, а программа разрабатывается в тесной связи и с ориентацией на эти признаки [5]. Тогда, зная заранее список таких параметров, потребитель запрашивает по ним систему МФРЯ и получает необходимые сведения с указанием, в какой работе, где, как и в каком аспекте рассматривается тот или иной вопрос. При необходимости расширить или углубить ответ он обращается в систему дополнительно или уже непосредственно к первоисточнику. Автор запроса сможет также быстро получить полную библиографию вопроса за определенный период.

Во-вторых, существует и другой путь отражения морфологии русского языка в МФРЯ. Грамматисты, лексикографы давно ощущают потребность в исчерпывающих сведениях о грамматическом употреблении конкретных слов в различных текстах разных хронологических периодов.

Создавая словари, мы часто не знаем, имеются ли какие-то морфологические ограничения в употреблении того или иного слова, а цитаты из

[5] Именно таким путем на кафедре математической лингвистики ЛГУ разработана автоматизированная система по исторической лексикологии.

картотек помогают не всегда. При этом в разных функциональных стилях, типах текстов морфологически слова ведут себя по-разному. Здесь МФРЯ создает идеальные предпосылки для реализации этой поистине грандиозной задачи — грамматической паспортизации русской лексики в зависимости от ее функционирования.

По самому своему замыслу такие фонды МФРЯ, как генеральный словник и иллюстрационно-текстовый фонд, должны содержать максимально исчерпывающие реестры слов, словоформ, контекстов для различных русских текстов. На основе этих фондов и может быть постепенно произведена всеобщая грамматическая паспортизация русской лексики, документированная реальными текстами. Наличие в будущем в МФРЯ исчерпывающих сведений по морфологии каждого слова трудно переоценить. Накопление таких фактов в МФРЯ резко повысит в перспективе частоту обращения к нему и экономическую эффективность его работы.

Таким образом, именно морфология, представляющая собой основной костяк, ядро системы любого из славянских языков, позволяет лучше увидеть целый ряд спорных лингвистических и информационных проблем проектирования и разработки МФРЯ, обсудить которые представляется целесообразным сейчас, на начальном этапе этой большой и сложной работы. Наконец, как информационная система МФРЯ не может не учитывать основных тенденций современной информатики, связанных, в частности, с проектированием не просто информационных систем, а систем экспертной оценки и классификации фактов [34—36]. Однако этот круг вопросов требует уже особого обсуждения.

ЛИТЕРАТУРА

1. *Андрющенко В. М.* Машинный фонд русского языка. Основные компоненты.— Уч. зап. Тартуского гос. ун-та, 1984, № 689.
2. *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы. Новосибирск, 1979.
3. *Ершов А. П.* Машинный фонд русского языка (Внешняя постановка вопроса).— ВЯ, 1985, № 2.
4. *Андрющенко В. М.* Машинный фонд русского языка: постановка задачи и практические шаги.— ВЯ, 1985, № 2.
5. *Богданов В. В., Бондарко Л. В., Буторов В. Д., Гердо А. С.* Теоретические проблемы языкознания и практические потребности народного хозяйства.— Вестник ЛГУ, 1983, № 8.
6. *Супрун А. Е.* Части речи в русском языке. М., 1971.
7. *Абакумов С. И.* Современный русский литературный язык. М., 1942.
8. *Виноградов В. В.* Русский язык (Грамматическое учение о слове). М., 1947.
9. *Буланин Л. Л.* Трудные вопросы морфологии. М., 1976.
10. *Шанский Н. М.* Очерки по русскому словообразованию. М., 1968.
11. *Лопатин В. В., Улуханов И. С.* Несколько спорных вопросов русской словообразовательной морфонологии.— ВЯ, 1974, № 3.
12. *Лопатин В. В.* Русская словообразовательная морфемика. М., 1977.
13. *Кубрякова Е. С.* Основы морфологического анализа. М., 1974.
14. Обзор работ по современному русскому литературному языку за 1974—1977 гг. Словообразование. М., 1982.
15. *Герд А. С.* О сегментации текста на морфологическом уровне.— В кн.: Теория языка. Методы его исследования и преподавания. Л., 1981.
16. *Герд А. С.* Семантика морфемы: значение или значимость? — В кн.: Структурная и прикладная лингвистика. Вып. 2. Л., 1983.
17. *Андреев Н. Д.* Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л., 1967.
18. Статистико-комбинаторное моделирование языков. М.— Л., 1965.
19. *Андреева Л. Д.* Статистико-комбинаторные типы словоизменения и разряды слов в русской морфологии. Л., 1969.
20. *Перцев Д. Г.* Лексико-статистические методы в автоматизации решения задач лингвистического обеспечения АИПС: Автореф. дис. на соискание уч. ст. канд. филол. наук. М., 1980.
21. *Пиотровский Р. Г.* Текст, машина, человек. Л., 1975.
22. *Бектаев К. Б.* Статистика речи: 1957—1972. Алма-Ата, 1972.
23. *Арзикулов Х. А., Сайчикова П. В.* Статистика речи: 1973—1985. Библиографический указатель. Самарканд, 1985.
24. *Городецкий Б. Ю.* Теоретические основы прикладной семантики: Автореф. дис. на соискание уч. ст. докт. филол. наук. М., 1978.
25. *Белоногов Г. Г., Губарь Н. Т., Новоселов А. П.* Морфологический анализ слов на основе словаря словоформы.— НТИ, сер. 2, 1975, № 9.

26. *Кобзарева Т. Ю., Лесскис Г. А.* Система автоматического морфологического членения текста.— НТИ, сер. 2, 1975. № 2.
27. *Белоногов Г. Г., Калинин Ю. П., Поздняк М. Ф., Хорошилов А. А., Яфаева Г. М.* Алгоритмы многоступенчатого морфологического анализа русских слов.— НТИ, сер. 2, 1983, № 1.
28. *Белоногов Г. Г., Кузнецов Б. А.* Языковые средства автоматизированных информационных систем. М., 1983.
29. *Ильин Г. М., Лейкина Б. М., Никитина Т. Н., Откупщикова М. И., Фитиалов С. Я.* Лингвистический подход к задаче построения информационной системы.— В кн.: Информационные вопросы семиотики, лингвистики и автоматического перевода. М., 1971.
30. *Шаляпина З. М., Афанасьева Л. А., Ельницкий Л. Л. и др.* Об англо-русском многоаспектном словаре с грамматическим обеспечением.— В кн.: Вопросы информационной теории и практики, 1975, № 27.
31. *Леонтьева Н. Н., Кудряшова И. М., Соколова Е. Г. и др.* Семантический компонент в системах автоматического понимания текстов. М., 1982.
32. *Марчук Ю. Н.* Проблемы машинного перевода. М., 1983.
33. *Пащенко Н. А., Кнорина Л. В., Молчанова Т. В., Чепиго Т. С., Шумилина А. Л., Яровенко О. И.* Проблемы автоматизации индексирования и реферирования.— В кн.: Информатика. Т. 7. М., 1983.
34. Тезисы докладов Всесоюзной конференции «Диалог „человек — ЭВМ“». Л., 1982.
35. Тезисы докладов республиканской научно-технической конференции «Интерактивные системы и их практическое применение». Кишинев, 1984.
36. Языки представления знаний и вопросы реализации экспертных систем. Владивосток, 1984.