УДК 001.103:002

ЗАДАЧИ СЛОВООБРАЗОВАНИЯ КАК СОСТАВНАЯ ЧАСТЬ ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ В ОБЛАСТИ ЕСТЕСТВЕННО-ЯЗЫКОВОГО ОБЩЕНИЯ

И. Е. Воронина

Воронежский государственный университет

Рассматривается задача компьютерного моделирования словообразовательных процессов и их диагностирования. Приводится пример вычислительного эксперимента с использованием диагностического аппарата.

Русское словообразование было объектом постоянного внимания исследователей на всем протяжении развития отечественного языкознания. Особенно возрос интерес к вопросам словообразования в последние десятилетия. Это непосредственно связано с общей интенсификацией изучения русского языка, выделением в нем разных уровней, с выявлением внутренних источников роста словарного состава языка в послеоктябрьский и послевоенный периоды, с расширением терминотворчества и необходимостью языкового регулирования, с вопросами машинного перевода и т. д. Как следствие, расширялась и углублялась проблематика, связанная с изучением словообразования, совершенствовались методы и принципы анализа, привлекались для исследования новые материалы [1]. Многие исследователи ставят перед собой чисто практические задачи: создать программу, которая моделирует поведение объекта [Н. Н. Перцова, А. В. Черемхин, А. В. Рафаева 2, 3, 4]. При этом никаких ограничений на способ решения задачи не накладывается. Использован предварительный список моделей словообразования, список словообразовательных морфем, морфов, дополнительные сведения, касающиеся построения основы требуемого вида по основе, заданной в словаре А. А. Зализняка. В рамках словообразовательного анализа для заданного входного слова ищется мотивирующее слово, совокупность его дериватов (тех слов, для которых оно является мотивирующим), строится словообразовательное гнездо. В рамках семантического анализа выполняются задачи коррекции в словообразовательном гнезде при несовпадении направления формальной и семантической деривации и приписывания толкования дериватам.

Работа И. Г. Милославского [5] посвящена словообразовательному синтезу. Задача этой работы состоит в том, чтобы, имея индекс морфем русского языка с их значениями и формальными преобразованиями, попытаться определить те правила, с помощью которых можно, во-первых, понимать членимые русские слова и, во-вторых, образовывать новые слова от данных по заданному семантическому различию (с. 4).

Е. А Карпиловская [6] ставит своей целью определение способов упорядочения глаголов в гнездах тематической группы звучания и разработка методики автоматизированного конструирования образуемых глаголами комплексов значений (на материале украинских глаголов звучания, образованные от звукоподражаний и существительных (1909 слов)). Решаются задачи определения способов формирования словообразовательной структуры глаголов, производных непосредственно от исходных слов гнезд и выявления формальной и семантической структуры глагольных комплексов; разрабатываются алгоритмы для автоматизированного конструирования комплексов глаголов по общим формальным и семантическим признакам. Что касается задачи сочетаемости морфем, то в работе Е. А. Карпиловской используются *реальные* окружения корней. Рассматривается не способ решения проблемы сочетаемости морфем, а проблема моделирования словообразования в выбранной подсистеме украинского языка.

Можно с уверенностью утверждать, что впереди еще большой и сложный путь к познанию системы русского словообразования как в целом, так и во многих ее отдельных звеньях.

Слово является важнейшей структурно-семантической единицей русского языка, которая служит для обозначения предметов, процессов,

свойств. В структурном отношении слово состоит из морфем. В словах закрепляются результаты познавательной деятельности, без слов невозможны не только выражение и передача понятий и представлений, но и само их оформление [7].

Р. Г. Пиотровский [80] утверждал, что поскольку естественный язык является важнейшим средством общения, следует предположить, что его единицы и связи могут быть обнаружены с помощью экспериментов и оценены путем измерений, использующихся в семиотике и теории информации. В его работе говорится о том, что статистические измерения информации в речи обычно осуществляются, исходя из следующих предположений:

1) речевое общение рассматривается как канал связи, по которому с помощью букв, звуков, морфем или других лингвистических единиц передается информация;

2) лингвистические единицы речи выступают в виде некоторого кода (языка); в коде-языке заданы ограничения не только на сочетаемость (комбинаторику), но и на вероятность появления этих единиц в речи;

3) соединенные речевым каналом источник и приемник, то есть говорящий и слушающий используют один и тот же код.

В [8] рассматривается эксперимент по угадыванию букв текста. При этом используется вероятностно-информационный математический аппарат, который дает среднюю условную энтропию (информацию) для каждой буквенной позиции, то есть возникает некоторая усредненная схема текста, в которой не отражена значимость отдельных лингвистических единиц (речь здесь идет о вероятностной оценке такой значимости).

Несмотря на то, что вышеупомянутая работа не затрагивает процесса словообразования, в ней содержатся важные выводы, которые способны повлиять на моделирование и диагностику данного процесса.

Так, например, стратегия угадчика показывает, как на комбинаторику фигур — букв и слогов накладываются ограничения с сочетаемостью простых знаков — морфем. В свою очередь комбинаторика морфем ограничивается сочетаемостью знаков более высокого порядка — слов. Затем по мере развертывания текста на комбинаторику слов накладываются ограничения в сочетаемости словосочетаний и предложений, а на это последнее, в свою очередь, — экстралингвистические композиционно-сюжетные ограничения всего повествования.

Тем самым, по сути дела, в [8] показывается, в каком направлении можно укрупнять лингвистические объекты, моделируя и изучая естественно-языковое общение.

Интересным представляется также вывод о том, что морфемные схемы внетекстовых слов в трех, взятых для исследования языках, характеризуются последовательным чередованием максимумов и минимумов информации. При этом максимумы информации попадают на первую, а минимумы на последнюю букву морфемы.

Рассматривая слово в качестве лингвистического объекта, можно выдвинуть гипотезу о правилах его образования, моделируя этот процесс на компьютере. Функцию обратной связи в этом случае может выполнять диагностика словообразования, которая позволила бы устанавливать степень соответствия (приближения) модели реальному словообразовательному процессу.

Представляется целесообразным представить процесс словообразования при помощи познавательной модели. Познавательная деятельность ориентирована в основном на приближение модели к реальности, которую модель отображает.

С применением компьютеров в лингвистических исследованиях задача поиска системы и единых принципов ее организации приобрела особую актуальность. Гипотезы, основанные на единстве, доказали свою продуктивность. Так И. В. Гете высказывалось предположение о том, что скелеты всех животных являются реализациями одного идеального скелета, который в [9] назван типом, а все части растений являются видоизменениями листа. В. Я. Пропп обнаружил, что все волшебные сказки по своей структуре являются вариантами одной сказки с единой структурой [10]. В [11] в предположении единства этапов человеческой цивилизации, дается сравнительный анализ человеческих обществ. Н. И. Вавилов добился замечательных результатов, предположив единство варьирования генетически близких растений [12].

На фоне этих примеров будет естественным предположить, что все слова русского языка являются реализацией одного идеального слова (ПРАслова, как его назвал бы И. В. Гете), точнее одного принципа образования слов.

При отсутствии эксплицитных формулировок правил порождения русского слова, наличие некоторых закономерностей, которым подчиняется словообразовательный процесс, делает возможным попытку его формализации и обобщения. К таким закономерностям относится морфемная структура слова, регулярный характер присоединения приставок и суффиксов, обязательное наличие корня и окончания.

Наличие определенных правил порождения русского слова ставит две взаимосвязанные задачи:

(1) формализации и программного подтверждения этих правил;

(2) фиксации новых правил.

Разработанная модель не является отображением словообразовательного процесса «один к одному». Она скорее является инструментом изучения русского словообразования. Если полагать, что *адекватная* модель — это та модель, с помощью которой достигается поставленная цель, то введенное понятие адекватности не полностью совпадает с требованиями полноты, точности и правильности. В этом случае адекватность означает, что эти требования выполнены не вообще, а лишь в той мере, которая достаточна для достижения цели. *Цель — это ответ на вопрос, почему генерируемое слово является русским словом.* Для достижения цели необходимо построить систему правил, по которым формируется слово. Наличие правил позволит исключить из рассмотрения так называемый отрицательный материал. Отсечение отрицательного материала происходит с помощью фильтрации. Процесс фильтрации осуществляется с помощью системы фильтров, которые имеют вид правил, реализующих запреты на определенные сочетания структурных составляющих слова — морфем. Набор фильтров не закрыт. Он предполагает в добавление новых фильтров в их систему, которая имеет иерархическую структуру. Программа предоставляет средства, необходимые для получения и накопления словообразовательного материала и его анализа, для того чтобы сформулировать правила нового, более высокого уровня, для их проверки и подтверждения. Поэтому модель является расширяемой. Она спроектирована таким образом, что подразумевает свое расширение в процессе эксплуатации (этот процесс можно назвать эволюцией модели) добавлением новых компонентов, в виде фильтра очередного уровня. Конечным этапом эволюции является полнота набора фильтров, что и приведет к главной цели модели — будут сформулированы и программно реализованы правила построения русского слова.

Для того, чтобы оценить степень расхождения модели с реальностью необходим инструмент диагностирования. Для создания такого инструмента можно применить вероятностно-статистические методы теории информации), в частности, энтропийную модель. Диагностический инструментарий является внешним по отношению к модели словообразования.

Таким образом, модель словообразовательного процесса играет роль инструмента для накопления, анализа, изучения материала, выполняя функцию обучения, а набор диагностических средств используется в исследовательских работах, осуществляя функцию оценки.

Следует отметить, что на этом функция диагностических средств не исчерпывается. С их помощью можно описать поведение словообразовательной системы, диагностируя каждый этап в создании слова как такового. В этом смысле можно рассматривать эти средства как модель поведения системы при ее переходе из одного состояния в другое, по мере выбора и добавления структурных составляющих слова (морфем) до его полного формирования. Задав направление генерации слова, выбрав его вид (формулу), имея в наличии правила, описывающие формирование слова, при разработанной на момент исследования системе фильтров, которые составляют основу для диагностирования, можно выявлять закономерности словообразовательного процесса, поставить вопросы о причинах поведения словообразовательной системы тем или иным образом при переходе из одного состояния в другое, тем самым получая новые знания. Схематично этот процесс представлен на рис 1.

Для поддержки диагностики разработаны алгоритмы, получившие программную реализацию.

Рассматривая слово как знак, мы тем самым подразумеваем наличие в нем двух составляющих: означающего и означаемого. Означающее — это материальная оболочка, план выражения слова. Означаемое — это план содержания, значение слова. Означающее называют лексемой, а означаемое (одно из значений слова) — семемой. В работе речь идет о синтезе

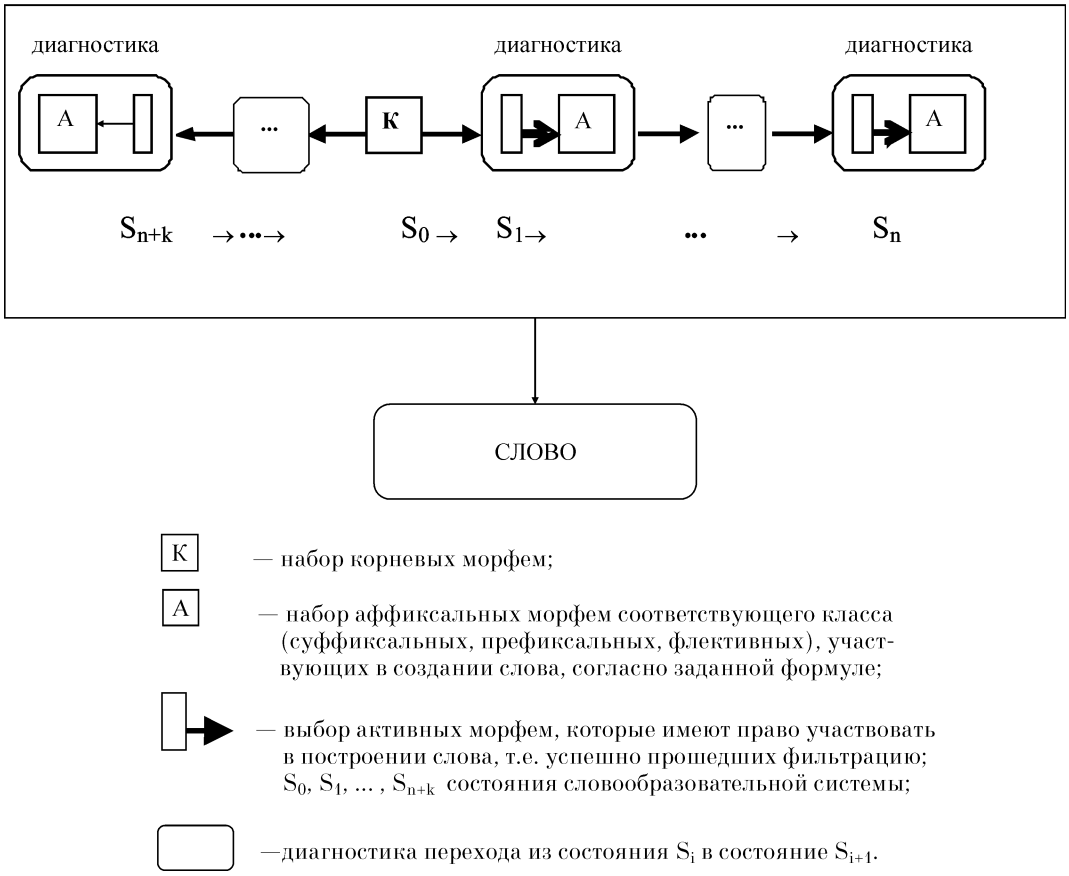

Рис. 1. Динамика исследовательского процесса

лексем. К синтезу осмысленных оболочек слова мы идем формальным путем.

Сначала мы работаем с отрицательным материалом. Под отрицательным материалом понимается последовательность морфем, не являющаяся словом. Базой для синтеза отрицательного материала является инвентарь морфем. Синтез происходит по выбранной исследователем формуле. Наполнение составляющих формулы происходит двумя путями. Первый способ заключается в полном переборе имеющихся морфем, что подразумевает рассмотрение всех возможных комбинаций морфем при заданном способе их сочетания. Второй способ связан с заполнением составляющих формулы случайным образом.

Любая отрицательная последовательность морфем заключает в себе проблему, почему она не является словом. Ответ на такой вопрос предполагает выход к новому знанию.

Можно задавать направление процесса синтеза. Для этого предусмотрена возможность фиксации отдельных составляющих формулы. Кроме того, имеющееся в наличии множество морфем может сужаться. Сужение исходного множества может происходить прямым выбором и выделением активных морфем, а также заданием списка активных морфем с помощью маски.

Поиск ответа на вопрос о том, почему не всякую последовательность морфем можно считать словом, приводит к формулировке правила. Правило формулируется в форме запрета. Программная реализация запретов на сочетаемость морфем приводит к созданию языковых фильтров.

Последовательность морфем, пропущенная первым фильтром, является положительным материалом с точки зрения данного фильтра, но с точки зрения осмысленности этих последовательностей морфем, они могут относиться как к положительному, так и к отрицательному материалу, причем отрицательный материал преобладает. Анализ этого отрицательного материала (назовем его отрицательным материалом второй степени) приводит к формулировке нового правила (системы правил) и, соответственно, к созданию нового фильтра. Процесс (в идеале)

продолжается до тех пор, пока созданная таким образом иерархия фильтров не начинает порождать только осмысленные последовательности морфем.

Для проведения лингвистических исследований в области словообразования А. А. Кретовым в [12, 13] предлагается и обосновывается введение особой транскрипции, которая получила название *этимологическая*.

Например:

внимательнейший	{вън-ьм=А=тел=ьн=ейсй_ий}
влияние	{въ-лиj=А=ній_е}
отвлеченность	{от-велк= Ø =Ен=ьн=ост_ь}

В результате мы работаем с синтезированным или расчлененным словом не в его орфографической форме, а в этимологической транскрипции, которая дает глубинный уровень представления слова.

Этимологическая транскрипция позволяет значительно уточнить наши представления о количестве и качестве морфем и их сочетаемости.

Создание языковых фильтров — ключевая задача словообразования. Применение фильтров носит иерархический характер. Сначала сгенерированное слово пропускается через фильтр самого нижнего уровня, затем последовательно применяются фильтры следующих уровней, в случае успешного результата предыдущего уровня. При этом для исследователя представляют интерес количественные характеристики: общее число сгенерированных слов, количество слов, прошедших каждый фильтр и количество сгенерированного положительного материала. По сути дела, строится частная модель, моделирующая такое языковое явление, как синтез лексем. Базовым множеством для построения модели является инвентарь морфем русского языка. Результатом функционирования модели является последовательность морфем («слово»). Фонетический фильтр является фильтром самого нижнего уровня. Правила фонетического фильтра касаются суффиксального стыка морфем, а именно, налагается ряд запретов на сочетаемость гласных и согласных. Дередупликативный фильтр накладывает запреты на удвоение аффиксов и приставок. Кроме того, на текущий момент разработаны еще два типа фильтров — флективный и частеречный.

Анализируя отрицательный материал, мы самообучаемся, осознаем то, чем интуитивно пользуемся, то есть языковые знания, и формализуем настолько, чтобы этим можно было пользоваться.

Применение каждого фильтра создает предпосылки для создания последующих. Анализ результатов состоит в оценке положительного материала N-ой степени с точки зрения осмысленности порождаемой последовательности морфем.

Анализируя, мы пытаемся разделить понятия положительного материала на относительный положительный материал N-ой степени и абсолютный положительный материал, обладающий свойствами осмысленности.

Рассмотрим конкретный пример применения диагностических средств.

В качестве исходного материала используем «Морфемно-морфонологический словарь языка А. С. Пушкина» (около 23 000 слов) [14]. Материал представлен в следующем виде: слова разбиты на морфемы, для каждого слова указана частота его употребления в тексте. Тем самым можно получить наборы морфем, задействованные при образовании слов вышеупомянутого словаря, а так же выявить частоты их употребления. Известна реальная сочетаемость морфем.

Была просчитана полная энтропия, то есть энтропия без учета фильтрации, экспериментальная энтропия, то есть энтропия с использованием имеющейся системы фильтров, а так же реальная энтропия, то есть использованы сведения о реальном сочетании морфем при образовании слов словаря. Результаты были получены на основании цепей Маркова 1-го порядка.

Поскольку, как упоминалось ранее, разработанные фильтры, практически не затрагивают префиксные стыки, представлялось интересным получить результаты для слов, образованных по следующим формулам:

К-О, К-С-О, П-К-О, К-С-С-О, П-К-С-О,
П-П-К-О, К-С-С-С-О,

где К — корень, С — суффикс, О — окончание.

В таблице приведены результаты вычислений реальной, полной и экспериментальной энтропии. Нас будет интересовать отклонение экспериментальных результатов от реальных. По сути дела, по этому отклонению можно будет судить о степени приближения полученных результатов к реальности или говорить об *эффективности фильтрации*.

Эффективность фильтрации при наличии набора вышеупомянутых фильтров представлена на рис. 2.

Таблица

Результаты вычислительного эксперимента

Номер формулы (N)	Реальная энтропия (РЭ)	Экспери-ментальная энтропия (ЭЭ)	Полная энтропия (ПЭ)
1)	11,592	11,544	12,912
2)	13,552	21,249	29,897
3)	5,690	19,671	31,667
4)	14,176	22,809	48,980
5)	8,170	23,650	42,038
6)	1	18,518	30,449
7)	15,348		66,227

Проведенный вычислительный эксперимент продемонстрировал возможность применения энтропийной модели для диагностики процесса словообразования. Использованный математический аппарат может применяться для того, чтобы при разработке и применении новых фильтров можно было оценить их действенность, то есть степень приближения к реальной ситуации, когда на выходе получается то, что мы считаем русским словом. Кроме того, при развитой системе фильтров можно оценить «вес» каждого фильтра, что приведет, в свою очередь, к выяснению степени влияния тех или иных правил на процесс порождения слова и, следовательно, к выявлению тенденций, имеющих место в словообразовательном процессе. Другой стороной вышеописанных действий можно считать поиск ответа на вопрос, какая последовательность морфем является словом и почему.

СПИСОК ЛИТЕРАТУРЫ

1. *Максимов В.И.* Структура и членение слова. Ленинград: Изд-во Ленинградского университета, 1977. — 148 с.

2. *Перцова Н.Н., Черемхин А.В.* Эксперименты по построению формальной модели русского словообразования // Труды машинного фонда русского языка. Т. 2. М.: ИРЯ РАН, 1992. С. 86—103.

3. *Рафаева А.В.* Система РУСЛО-2: автоматический анализ и синтез производных слов русского языка // Проблемы компьютерной лингвистики: Сб. научн. трудов / Под. ред. А. А. Кретова. — Вып. 1. — Воронеж: РИЦ ЕФ ВГУ, 2004. С. 96—102.

4. *Percova N.* RUSLO: An Automatic System for Derivation in Russian // Amsterdam/Philadelphia, 1996.

5. *Милославский И.Г.* Вопросы словообразовательного синтеза. М.: Изд-во Московского ун-та, 1980, — 296 с.

6. *Карпиловская Е.А.* Конструирование глагольных зон словообразовательных гнезд (на материале украинских глаголов звучания): автореф. дис. 10.02.21 — Структурная, прикладная и математическая лингвистика, канд. филол. наук. Л., 1987. — 18 с.

7. *Шемакин Ю.И.* Начала компьютерной лингвистики. М.: Изд-во МГОУ, А/О «Росвузнаука», 1992. — 114 с.

8. *Пиотровский Р.Г.* Информационные измерения языка. Л.: Наука, 1968. — 115 с.

9. *Гете И.В.* Избранные сочинения по естествознанию. Л.: Изд-во АН СССР, 1957.

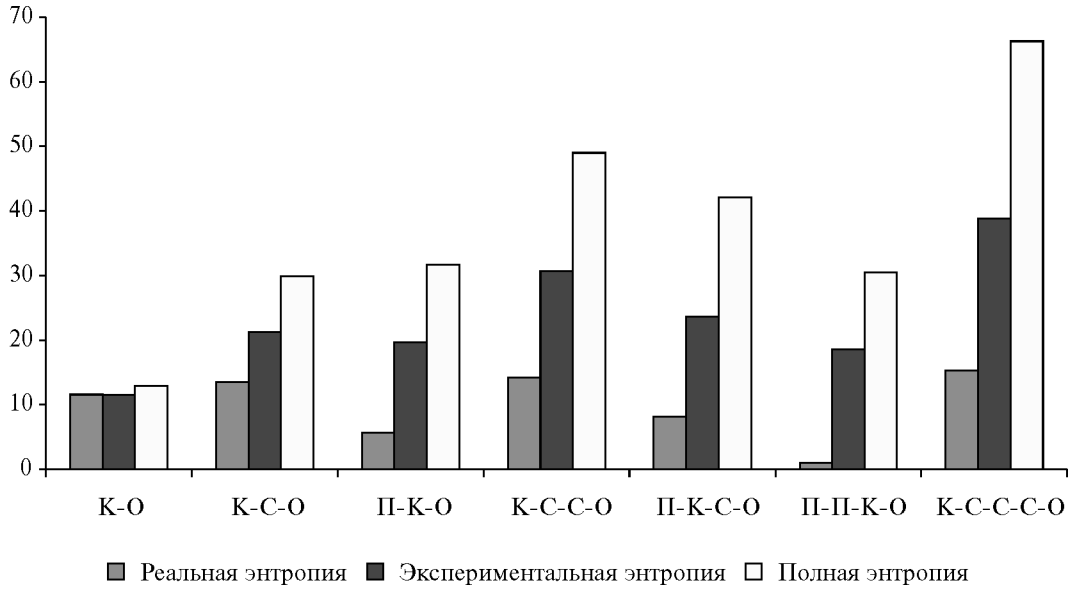

Рис. 2. Эффективность фильтрации

10. *Преображенский А.Б.* Состояние развития современных ЕЯ-систем // Искусственный интеллект: В 3 кн. Кн.1. Системы общения и экспертные системы: Справочник / Под ред. Э. В. Попова. М.: Радио и связь, 1990. С. 32—94.

11. *Шпенглер О.* Закат Европы. — Новосибирск: ВО Наука, Сибирская издательская фирма, 1993. — 592 с.

12. *Кретов А.А.* Фонема: аксиоматика и выводы // Вестник ВГУ. Серия лингвистика и межкультурная коммуникация, 2001, № 2, С. 50—63.

13. *Кретов А.А.* Теоретические и практические аспекты создания морфемного словаря // Вестник ВГУ. Серия лингвистика и межкультурная коммуникация, 2002, № 2, С. 55—61.

14. *Кретов А.А., Матыцина Л.Н.* Морфемно-морфонологический словарь языка А. С. Пушкина: Ок. 23000 слов. Воронеж: Центрально-Черноземное книжное издательство, 1999. — 208 с.