Компьютерное моделирование лингвистических объектов

Воронина И.Е.

Взаимодействие человека и ЭВМ усиливает прикладное начало в языкознании. Бодуэн де Куртенэ, говоря о задачах языкознания, которые необходимо решить в 20 веке, отмечал: " Нужно чаще применять в языкознании количественное, математическое мышление и таким образом приблизить его все более к наукам точным" [1].

Прикладные лингвистические задачи отличает их заказной характер. В большинстве своем они представляют собой тот или иной социальный заказ. Их реализация протекает в диалоге "заказчик-разработчик". Второй особенностью прикладных задач является их проверяемость, при этом проверяемость повторная, неоднократная и каждый раз на новом материале.

Язык - семиотическая система с иерархической структурой. Эта система нуждается в изучении и формализации, что и делает разработку, реализацию и унификацию подходов и методов ее познания крайне актуальной. Под унификацией понимается применение для изучения каждого уровня одной и той же совокупности методов.

Формализация естественного языка является нетривиальной задачей и обладает всеми особенностями слабоструктурированных проблем.

Представляется разумным использовать тот факт, что язык - открытая система закрытых подсистем. Каждая подсистема конечна, следовательно, ее можно моделировать, а затем устанавливать определенные отношения между подсистемами.

Будем рассматривать выявление правил синтеза, то есть формализацию порождения правильных цепочек на заданном языковом уровне. Предлагается формировать правила в виде запретов на сочетаемость для каждого языкового уровня.

Практически формулировать правила могут только эксперты в предметной области. Но для того, чтобы эти правила были по-настоящему формализованы, необходимо ограничить способ подачи таких правил определенными рамками. Следующий уровень формализации - математический, программный и т. д. Таким образом, материал для описания правил строится на базе коммуникативных процессов, то есть в конечном итоге это вопросы, сформулированные в терминах, понятных эксперту, но форма подачи ответа жестко оговаривается. Правила формулируются в виде $a \rightarrow b$ (ЕСЛИ a, ТО b), где b трактуется всегда одинаково - запрет на сочетаемость (отрицательный вердикт). Следовательно, наибольший интерес представляет именно условие a, которое собственно и определяет правило.

Выявленные и программно подтвержденные правила являются основой фильтров, позволяющих отсекать так называемый отрицательный материал, порождение которого неизбежно.

Модель языковой подсистемы является частью исследовательского инструмента, включающего в себя определение исходной подсистемы и сбор данных. Обработка данных определяет поведение модели. Лицо, принимающее решение (ЛПР), занимается интерпретацией данных, именно его решение влияет на пополнение модели (добавление новых признаков, новых правил, определяющих запреты на порождение определенных цепочек). Но ЛПР (эксперту, исследователю) необходим инструмент для обеспечения обоснования и поддержки принимаемых решений.

Диагностика процесса и будет инструментом, обеспечивающим принятие решения. Концептуальная схема проведения исследований естественного языка приведена на рис. 1.

Под объектом (рис.1) понимается объект моделирования, то есть языковые процессы определенного уровня иерархии. Моделирование объекта происходит на основании стартовой информации, делающей возможной саму попытку начальной формализации и,

следовательно, автоматизации. Поскольку речь идет о порождающей модели, результатом ее функционирования будет сгенерированный материал, подлежащий наблюдению и изучению.

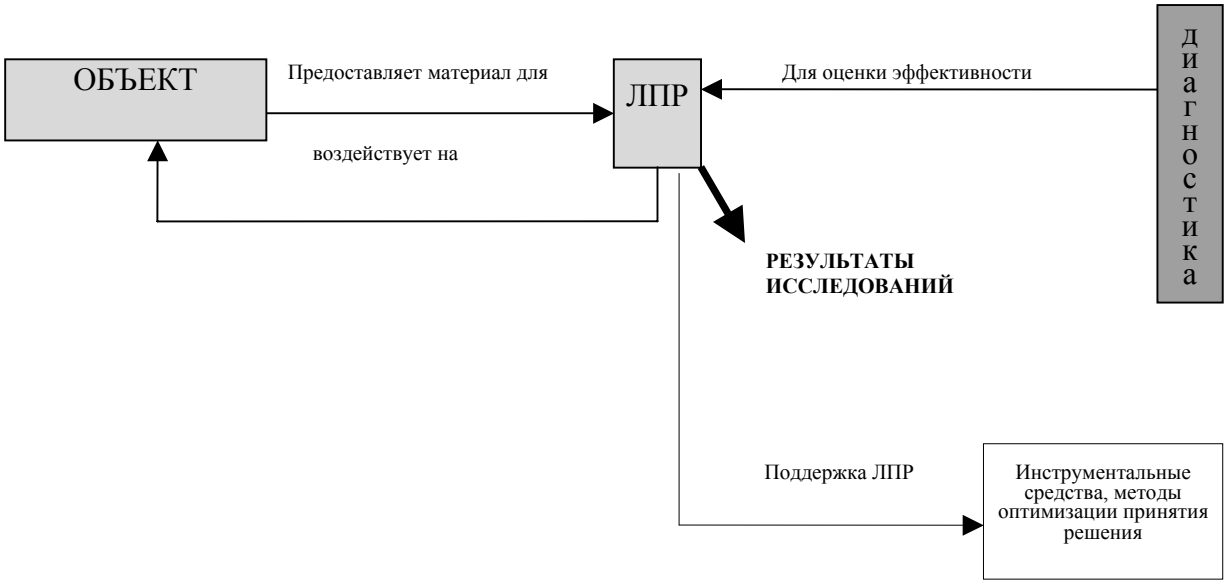

Рис. 1 Концептуальная схема проведения исследований

Использование диагностики уместно на этапе создания фильтра очередного уровня. При включении его в существующую иерархию фильтров можно продиагностировать наличие отклонения модели от реальной системы (подсистемы) и получить необходимые количественные оценки. Этапы проведения исследований приведены на рис. 2.

Следует заметить, что, когда идет речь о фильтре очередного уровня, предполагается наличие иерархической системы фильтров для исследуемой и формализуемой языковой подсистемы. Такая система фильтров строго ориентирована на исследуемую подсистему и отражает ее специфику. Но для каждой такой подсистемы можно использовать и метафильтр, при создании которого применяется универсальный математический аппарат, опирающийся на методы детерминационного анализа и методы принятия решения в недостаточно определенной ситуации [2-4].

Проиллюстрируем вышеизложеннные соображения на примере решения словообразовательных задач [5-7].

В качестве основного подхода к вопросам автоматизации словообразовательного процесса принят метод лингвистического эксперимента, предложенный И.А. Бодуэном де Куртенэ в 19 веке и методически обоснованный Л.В. Щербой в начале 20 века. Согласно этому методу правила словообразования диагностируются посредством обращения к отрицательному материалу, который анализируется и служит базой для выявления, формализации и подтверждения правил порождения слов. Достоинства отрицательного материала трудно переоценить, поскольку он позволяет ставить, а затем и решать вопросы, которые не возникают и не могут возникнуть при, по существу, автоматическом пользовании положительным материалом родного языка. Синтез отрицательного материала осуществляется из реально существующих морфем, занимающих позиции, отвлеченные от реально существующих слов.

Несмотря на отсутствие точных формулировок порождения русского слова, процесс словообразования подчиняется некоторым закономерностям, дающим право попытки обобщения и формализации. В частности, слово состоит из морфем. Известно

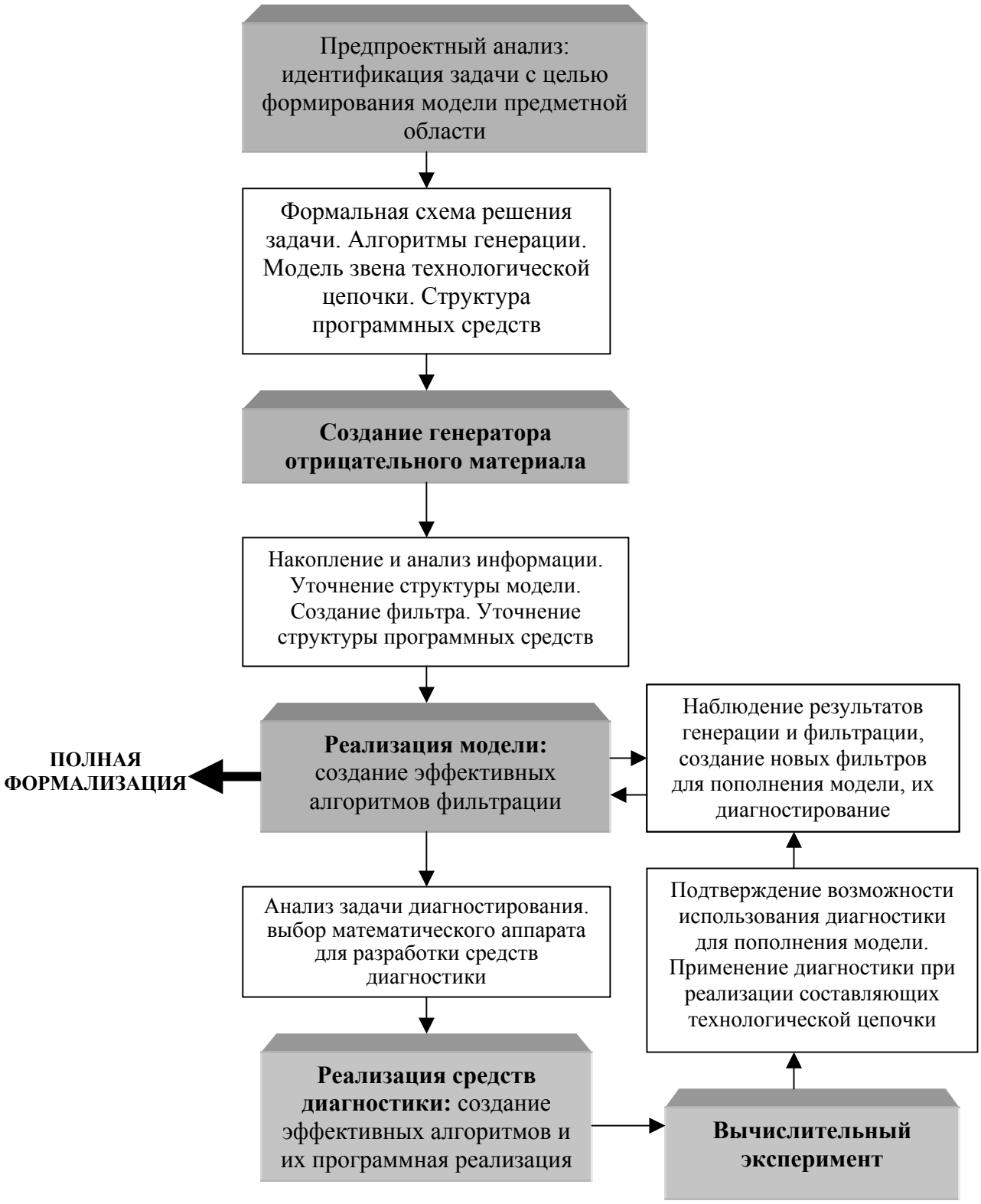

Рис. 2 Этапы проведения исследований

также, что в слове всегда присутствует корень и окончание (окончание может быть нулевым), а присоединение приставок и суффиксов носит регулярный характер. Кроме того, количество морфем конечно (имеются ввиду исконно русские, а не заимствованные из других языков морфемы), и их число достаточно невелико. Так, например, продуктивных корней не более 3000, суффиксов и приставок не более 500, а число окончаний (для словарных форм) не превышает 50. На процесс порождения слова накладываются следующее ограничения: слово не должно содержать не более 7±2 морфем, не считая окончания (рассматриваются только слова с одним корнем). Следовательно, предельная формула русского слова - **П3-П2-П1-К-С1-С2-С3-С4-С5-С6-Ф**, где П - приставка, К - корень, С - суффикс, Ф - окончание (флексия). Перечисленные закономерности служат хорошей базой для автоматизации моделирования порождения русского слова, когда роль генератора отрицательного материала берет на себя компьютер.

Отрицательный материал служит основой для выявления правил построения русского слова. В свою очередь, наличие правил порождения слова позволяет поставить вышеупомянутые задачи формализации и программного подтверждения правил, а также фиксации новых правил.

Следует подчеркнуть тесную взаимосвязь этих двух задач. Дело в том, что четкие и однозначные формулировки правил порождения русского слова из имеющегося инвентаря морфем отсутствуют, поэтому фиксация всякого правила, в том числе и самого первого, требует программного подтверждения, проверки и - после накопления порождаемого материала и его анализа - становится базой для создания новых правил следующего (более высокого) уровня.

Необходимо заметить следующее. Если принять, что слово являет собой единство звучания и значения, то нами рассматривается только графическое порождение и отображение звуковых оболочек. Не ставится задача порождения слова по заданному значению и не осуществляется синтез значения слова (его содержательной стороны). Таким образом, программа порождения русских слов является, по сути, программой синтеза звуковых оболочек слов. Вместе с тем нельзя сказать, что при этом совершенно не учитывается семантический фактор, поскольку словом в программе считается не любая, а только осмысленная последовательность морфем, которая может быть связана (через понятия) с каким-то явлением внеязыковой действительности.

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 фильтрации 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. Дередупликативный фильтр стоит на втором уровне иерархии. Далее следуют флективный, частеречный и морфонологический фильтры.

После появления фильтра первого уровня и включения его в систему (как и для всех последующих фильтров), исследователь получает возможность проверить его действие, наблюдая и анализируя результаты. При удовлетворительном результате полученный материал становится базой для создания нового фильтра.

При включении каждого фильтра в систему происходит значительное сокращение перебора.

Так, например, в случае моделирования словообразовательного процесса только один лишь запрет на удвоение приставок сокращает число возможных вариантов сочетаний приставок следующим образом: для случая наличия двух приставок в формуле вместо n^2 вариантов, где n - количество приставок, задействованных в инвентарном наборе, мы имеем $n^2 - 1$ вариантов, а при наличии всех трех приставочных позиций - $(n-1)^2 * 3$, вместо n^3 вариантов. Правила жесткой связи корня с допустимыми окончаниями и аналогичное правило для суффиксов сужает валентность для корня до 2, а для суффикса до 3, вместо ожидаемого числа равного количеству всех окончаний, задействованных в инвентарном наборе. При этом отсечение осуществляется автоматически, что было бы достаточно тяжелым ручным трудом при отсутствии программных средств.

Если вышеперечисленные правила таковы, что можно вручную оценить количество возможных вариантов по сочетаемости, то это малопродуктивно и требует больших затрат времени при выведении валентности для сочетания суффикс-суффикс и корень-суффикс. В результате вычислительного эксперимента были получены цифры 57 и 30 соответственно (при задействовании 197 суффиксов). Но главным результатом, который может оказать помощь при проведении исследовательских работ в области словообразования, можно считать получение таблицы сочетаемости каждого корня с допустимым набором суффиксов и получение аналогичной таблицы для каждого суффикса. Разумеется, что при разработке и добавлении новых фильтров эти таблицы будут уточняться.

Программные средства позволяют также оценивать нагрузку, приходящуюся на каждый фильтр.

Очевидна трудоемкость вышеперечисленных действий. При использовании программных средств пользователь получает возможность более эффективно вести исследовательский процесс, существенно сокращая временные затраты на проведение упомянутых операций. Результаты диагностирования могут влиять не только на процесс принятия решений, но и играть роль при анализе полученного материала. Учитывая, что исследователем является специалист в области лингвистики, возможно, не владеющий математическим аппаратом (в данном случае были использованы вероятностно-статистические методы теории информации), создание диагностического инструмента, разработка самих методов диагностики предоставляет ему новые, дополнительные возможности исследования и формализации словообразовательного процесса. Практически исследователю предоставляются методы количественного оценивания направления процесса исследований в области словообразования, формализованный аппарат осуществления диагностирования.

В контексте вышесказанного уместно привести результаты вычислительного эксперимента на базе морфемно-морфонологического словаря языка А.С.Пушкина [8]. В данном случае работа велась с теми словами, которые удовлетворяли следующим условиям:

1. слово является существительным, прилагательным или глаголом;
2. слово заканчивается только заданными окончаниями.

Таких слов оказалось: **20291**.

На рис. 3 представлены результаты, которые иллюстрируют эффективность работы фильтров. Алгоритмы и математический аппарат проведения диагностики более подробно представлены в [5-7]. Реальная энтропия вычислена на материале словаря, то есть иллюстрирует реальную сочетаемость, а экспериментальная энтропия вычислена на базе инвентаря морфем, полученных из словаря [8] с учетом разработанной на текущий момент неполной системы фильтров. На горизонтальной оси отмечены типы формул, по которым было наиболее интересно получить результаты:

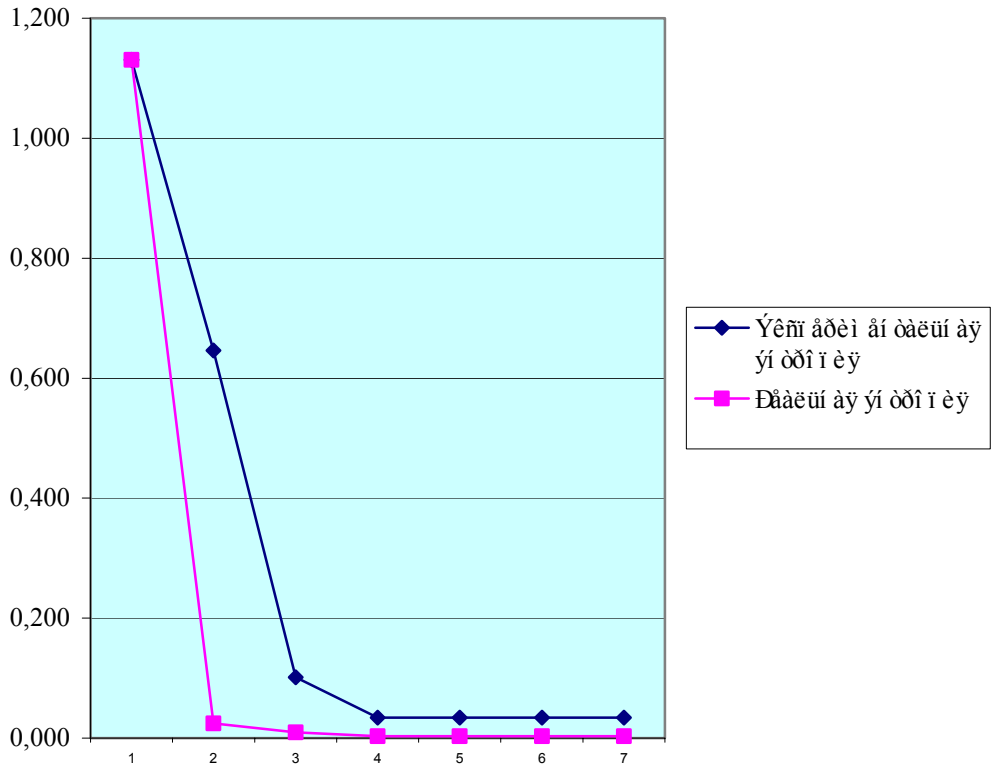

Рис. 3 Эффективность фильтрации на примере морфемно-морфонологического словаря языка А.С.Пушкина

1) К-Ф
2) К-С1-Ф
3) К-С1-С2-Ф
4) К-С1-С2-С3-Ф
5) К-С1-С2-С3-С4-Ф
6) К-С1-С2-С3-С4-С5-Ф
7) К-С1-С2-С3-С4-С5-С6-Ф.

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 до конца 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.

Глобальная цель – изучить структуру языка. Уровни структуры языка - это синтаксические предложения, слова, морфемы, фонемы. Все языковые уровни характеризуются наличием базовых элементов. Изучение языка может идти с двух позиций - анализа и синтеза, ибо выявленные правила синтеза могут способствовать проведению анализа, и наоборот. Для исследования и максимальной формализации каждой языковой подсистемы необходимо создавать программный инструментарий, реализующий процесс изучения путем выявления и проверки правил анализа и синтеза, тем самым максимально автоматизируя исследовательский процесс, освобождая при этом исследователя как от

рутинного процесса накопления и сбора информации, так и снимая вопрос трудоемкости ее обработки.

Литература

1. Бодуэн де Куртенэ И.А. Избранные труды по общему языкознанию. М.: Изд-во АН СССР, Т.2., 1963. 392 с.
2. Воронина И.Е. Детерминационный анализ как инструмент лингвистической онтогностики // Проблемы лингвистической прогностики: Сб. науч. трудов / Под ред. А.А. Кретова, Вып.2., Воронеж, 2002. С. 204-212.
3. Воронина И.Е. Метод "интуитивной оптимизации" в лингвистических исследованиях. // Математическое обеспечение ЭВМ. Межвуз. сб. научных трудов, Вып.3, Воронеж, 2001. С. 13-20.
4. Воронина И.Е. Принятие решения в случае недостаточно определенной сочетаемости структурных единиц при создании лингвистического обеспечения информационных процессов // Математическое обеспечение ЭВМ: Межвуз. сб. науч. тр., Вып. 2. Воронеж: Изд-во ВГУ, 2000. С. 21-25.
5. Воронина И.Е. Моделирование словообразовательных процессов // Тез. докл. 20й международной конференции "Системное моделирование социально-экономических процессов", Воронеж, 1998. С. 151.
6. Воронина И.Е. Проблемы формализации русского языка // Русский язык: исторические судьбы и современность: Международный конгресс исследователей русского языка (Москва, филол. фак. МГУ, 13-16 марта 2001): Труды и материалы / Под общей ред. М.Л. Ремневой и А.А. Поликарпова. М.: Изд-во МГУ, 2001. С.398-399.
7. Кретов А.А., Воронина И.Е. Лингвистическое обоснование программного синтеза слова (на материале русского языка) // Тез. докл. 2-й Международной конференции по квантитативной лингвистике "Qualico-94"- Москва, 1994. С. 187- 188.
8. Кретов А.А., Матыцина Л.Н. Морфемно-морфрнологический словарь языка А.А. Пушкина: Ок. 23 000 слов. Воронеж: Центрально-Черноземное книжное издательство, 1999. 208 с