

На правах рукописи

РАХИМБЕРДИЕВ Булат Нурланович

**ЭВОЛЮЦИЯ СЕМАНТИКИ
ЭКОНОМИЧЕСКОЙ ТЕРМИНОЛОГИИ
РУССКОГО ЯЗЫКА В XX ВЕКЕ**

Специальность 10.02.21 – Прикладная и математическая лингвистика

Автореферат

диссертации на соискание ученой степени
кандидата филологических наук

Москва 2003

Работа выполнена на кафедре прикладной и экспериментальной лингвистики Московского ордена дружбы народов государственного лингвистического университета

Научный руководитель — академик МАИ,
доктор филологических наук,
профессор
Марчук Юрий Николаевич

Официальные оппоненты — доктор филологических наук,
профессор
Городецкий Борис Юрьевич
— кандидат филологических наук,
доцент
Хроменков Павел Николаевич

Ведущая организация — Институт русского языка РАН

Защита состоится «_____» 2003 г. в _____ час. на заседании диссертационного совета Д 212.135.02 по защите диссертаций на соискание ученой степени доктора наук при Московском ордена дружбы народов государственном лингвистическом университете по адресу: 119992, г. Москва, ГСП-3, ул. Остоженка, д.38

С диссертацией можно ознакомиться в библиотеке Московского ордена дружбы народов государственного лингвистического университета.

Автореферат разослан «_____» 2003 г.

Ученый секретарь

В.С. Страхова

Данная работа посвящена поиску возможностей исследования исторического развития той части лексики русского языка, которая обслуживает экономические отношения общества.

Знание характера произошедших в языке сдвигов несет как прикладную, так и теоретическую ценность. Тогда как современная лексикография склоняется к пониманию языковой нормы как динамической, стихийно складывающейся структуры, именно ретроспективный характер языковой нормы обеспечивает ее условную стабильность. С вовлечением в экономический оборот широких кругов, ранее принимавших в нем минимальное участие, неизбежны стихийные новации и, наоборот, исключение прежних языковых средств из сферы коммуникации.

Исторические лексикографические исследования стремятся внести вклад в поддержание культурной преемственности нации, сохранение того, что было в какой-то момент достигнуто и накоплено. Однако в современных условиях для решения своих задач историческая лексикография располагает существенно меньшими ресурсами, чем лексикография синхронная, и тем более нуждается в высокопроизводительных инструментах исследования.

Таким образом, **актуальность** диссертации обусловлена вовлечением за последнее десятилетие в экономический оборот России значительного числа новых участников, что неизбежно ведет к сдвигам в составе и семантической структуре лексики этого подъязыка, представляющим несомненный интерес для лингвистического изучения. Ретроспективное исследование этого пласта лексики призвано зафиксировать и изучить его историю в XX веке в аспекте анализа соотношения языка и мышления на примере развития экономической лексики.

Настоящая работа преследует две последовательные **цели**: во-первых, нахождение наиболее эффективных методов извлечения данных об эволюции семантики экономической терминологии русского языка из имеющихся материалов XX века, и во-вторых, верификация и приложение этих методов к

исследуемому материалу для разработки диахронического словаря экономической лексики.

Для достижения указанных целей в рамках работы решаются следующие *задачи*:

- определение модели семантической структуры слова, обладающей необходимой объяснительной силой для анализа сдвигов лексической семантики на протяжении столетия, а также модели диахронических изменений семантики;
- оценка применимости существующих методов и подходов к извлечению данных о лексической семантике из различных источников к задаче диахронического анализа;
- разработка методики обнаружения изменений в содержании одной единицы плана выражения;
- разработка методики обнаружения диахронических расхождений в плане выражения одной единицы плана содержания (поиска диахронических синонимов), а также сопутствующего программного обеспечения;
- составление диахронического словаря экономической лексики;
- поиск диахронических синонимов в подязыке экономики.

Любая лексикографическая работа сопряжена с привлечением огромных ресурсов времени и труда лексикографов. Введение в лексикографическое исследование диахронического аспекта еще более увеличивает его трудоемкость. В этой связи основным *направлением* настоящего исследования стал поиск и создание инструментов, обеспечивающих высокую эффективность получения данных об экономической лексике русского языка.

Одной из тенденций современной лингвистики является представление данных о языке в лексикографической форме. Другой важной тенденцией является стремление к формализации лингвистических описаний [Караулов, 1986, 1982]. С одной стороны, стремление к формализации является следствием внутренней логики развития лингвистики, пришедшей к необходимости инвентаризации фактов языка и приведения собственной

методологии в соответствие с общенаучным требованием воспроизводимости результатов. С другой же стороны, только формализованный материал может быть обработан машинными методами, открывающими перед лингвистикой и лексикографией в частности совершенно новые возможности. Современный уровень развития вычислительной техники, открывший для лексикографии новую область корпусной лингвистики, позволяет извлекать обширные массивы нетривиальных языковых фактов из первичных текстов.

Из этих соображений вытекают *методологические установки* исследования:

- основной единицей изучения является слово, а основной формой представления практических результатов исследования – словарная;
- предпочтительными методами исследования являются такие, которые поддаются формализации и автоматизации, что, однако, не означает отказа от экспертного анализа там, где он более эффективен;
- из всего многообразия машинных методов в настоящем исследовании приоритет отдается методам корпусной лингвистики.

Научная новизна диссертации заключается в том, что:

- впервые обоснована целесообразность и возможность использования толковых словарей для анализа изменений лексической семантики русского языка в XX веке;
- уточнены и дополнены принципы и методы извлечения информации о сигнификативном и коннотативном уровнях семантики слова из статьи толкового словаря;
- впервые разработана методика автоматического извлечения диахронических синонимов из параллельных переводов иноязычного текста;
- разработаны новые приемы автоматической обработки текста на русском языке в дореволюционной орфографии и с высоким уровнем дефектов распознавания.

Теоретическая значимость диссертации заключается в том, что выполненный анализ современных представлений о структуре семантики слова и ее эволюции, а также методов их исследования, позволил определить пути эффективного изучения исторической семантики лексики русского языка в ретроспективе XX века, заключающиеся в экспертном анализе дефиниций генетически связанных толковых словарей, а также статистическом машинном поиске диахронических синонимов по переводам иноязычных текстов, выполненным в разные годы.

Практическая значимость диссертации заключается в том, что составленный исторический словарь экономической лексики русского языка XX века (объем 210 единиц) и данные по обнаруженным диахроническим синонимам в подязыке экономики (65 пар) могут использоваться при составлении различных словарей по экономическому слою русского языка, а также при чтении курса основ экономической теории, особенно в языковых вузах, и преподавании русского языка. Полученные методики сопоставления определений толковых словарей в целях извлечения данных о семантических сдвигах, а также разработанное программное обеспечение для автоматического извлечения диахронических синонимов из параллельных текстов на русском языке, в частности, в дореволюционной орфографии, могут использоваться для проведения диахронических исследований семантики любых других сфер русского языка. Созданный корпус машиночитаемого текста в дореволюционной орфографии представляет интерес для исторической лингвистики.

Достоверность и обоснованность выводов обеспечивается теоретически оправданным отбором материала, а также его объемом: при анализе словарных определений было произведено полное обследование всех экономических единиц, входящих в тезаурус по экономике и демографии и используемые словари; при анализе параллельных текстов использовались массивы объемом около 300 тысяч и 500 тысяч знаков.

В рамках исследования проведены два эксперимента, в каждом из которых используется соответствующая *методика и материал*. Основой первого эксперимента послужила теория диахронического вектора слова М.В. Марчук, перенесенная на материал толковых словарей русского языка и дополненная представлением о многоуровневой структуре семантического значения слова. Экономическая лексика, выделенная из всего объема общеупотребительной лексики литературного русского языка с помощью «Информационно-поискового тезауруса ИНИОН по экономике и демографии» (2001 г.), была проанализирована по данным «Толкового словаря русского языка» Д.Н. Ушакова 1935 – 1940 г.г. издания и «Словаря русского языка» С.И. Ожегова 1991 г. издания. Второй эксперимент выполнен в русле корпусных исследований и заключается в обработке двух русскоязычных переводов «Исследования о природе и причинах богатства народов» А. Смита (1895 и 1931 годов) набором вычислительных средств, близких к инструментам статистического машинного перевода, с целью извлечения набора диахронических синонимов.

На защиту выносятся следующие положения:

- На современном уровне развития лингвистики наиболее эффективным инструментом исследования исторического развития семантики определенного слова русского языка в ретроспективе XX века является экспертный анализ определений соответствующего слова в толковых словарях одной традиции разных лет;
- Сопоставление словарных определений одной лексической единицы в толковых словарях русского языка различных лет издания позволяют установить сдвиги на сигнификативном, денотативном и коннотативном уровнях семантики этой единицы;
- Для исследования исторического развития формы определенного означаемого (поиска диахронических синонимов) возможно применение методов анализа параллельных текстов к различным переводам одного иноязычного текста.

Апробация материалов и результатов исследования:

Основные положения и результаты настоящей работы нашли отражение в трех печатных публикациях, а также представлялись на следующих конференциях:

1. На X сессии Российского акустического общества, Москва, 2000
2. На международном конгрессе «Русский язык: исторические судьбы и современность», Москва, 2001 г.

Отдельные положения работы обсуждались на заседаниях кафедры прикладной и экспериментальной лингвистики МГЛУ.

Диссертация состоит из введения, трех глав, заключения, списка использованной литературы и трех приложений. Во введении обосновывается выбор темы исследования, характеризуется его актуальность и новизна, а также направление и методы исследования, формулируются цели и задачи, излагаются положения, выносимые на защиту.

Первая глава диссертации посвящена теоретическому анализу современных представлений о семантической структуре слова с целью выбора рабочей модели самой структуры и ее эволюции.

Во второй главе рассматриваются подходы, позволяющие выполнять высокопроизводительный анализ эволюции лексической семантики. Предлагаются два подхода: сопоставление словарных толкований одного слова в генетически связанных словарях разных лет с целью определения сдвигов означаемого одного означающего и вычислительное сравнение текстов разных лет с целью обнаружения изменения означающего одного означаемого.

В третьей главе диссертации описывается постановка и результаты двух экспериментов по применению указанных подходов на материале экономической лексики. В результате сопоставительного анализа определений 210 единиц экономической лексики в словаре Ушакова и словаре Ожегова 1991 г. издания определяются фактические возможности и ограничения данного подхода, а также формируется словарь семантических сдвигов

экономической лексики за соответствующий период. Эксперимент по созданию и применению вычислительной методики поиска диахронических синонимов, основанной на современных работах по статистическому машинному переводу, проводится на материале двух переводов «Исследований» Адама Смита разных лет. Помимо практического результата – списка обнаруженных диахронических синонимов и набора программных инструментов для их поиска – данный эксперимент позволяет дать оценку применимости корпусных методов в диахронической лексикографии.

В заключении формулируются выводы, а также указываются возможности практического использования результатов исследования.

Список литературы состоит из 106 наименований источников, использованных при работе над диссертацией.

В приложениях приводятся: схема изменений в семантике экономической лексики по данным толковых словарей, алгоритмы, созданные и использованные для поиска диахронических синонимов в параллельных текстах, а также образец выравнивания текстов 1895 и 1931 г.г. – результат одной из важных стадий обработки параллельных текстов.

В современной лингвистике значение слова рассматривается как сложная многоуровневая структура. В целях данного исследования используется четырехчастная система вертикальной связи словесного знака с внеязыковой действительностью:

- означающее (материальная, то есть письменная или звуковая сторона знака);
- сигнификат, соответствующий мыслительному уровню понятий и понимаемый как набор признаков класса объектов;
- денотат, понимаемый как класс объектов внеязыковой действительности, ограничиваемый сигнификатом (для слов абстрактной семантики может быть неотличим от сигнификата);

- референт (единичный предмет или ситуация, на которую указывает означаемое слова в конкретном акте речи).

В качестве горизонтальной модели лексической семантики принимается представление о возможности членения отдельного значения слова на семы, состоящие друг с другом в иерархических отношениях. Такое представление о структуре семантического значения позволяет говорить о родо-видовой, гипонимической, тематической и иной соотнесенности слова, определяемой набором его дифференцирующих семантических признаков.

Эволюция отношений номинации, рассматриваемая с точки зрения семантики, распадается на два явления. Изменение означаемого одного означющего представляет собой явление, известное как лексико-семантическое варьирование. Изменению означаемого одного концепта до сих пор уделялось меньше внимания, чем первому явлению. Часть работы посвящена именно исследованию диахронического варьирования плана выражения.

Лексико-семантическое варьирование представляет собой тот процесс, синхронным срезом которого является полисемия. Этот факт дает возможность рассматривать методы и инструменты исследования лексической полисемии как потенциальный источник средств исследования исторических лексико-семантических изменений. Современные представления терминоведения, допускающие не только диахроническую вариативность семантики термина, но и его полисемию, не накладывают дополнительных ограничений на выбор инструментов в связи с терминологическим характером изучаемого материала [Татаринов, 1996].

Методы исследования лексической полисемии условно делятся на контекстологические, структурные, психолингвистические и корпусные подходы. Применимость психолингвистических методов в исторических исследованиях весьма ограничена. Понятие корпусных методов фактически определяет подход к объему и характеру первичного материала, и с точки зрения методологии исследования может подразумевать как

контекстологические, так и психолингвистические работы. Заметим также, что в настоящее время корпусные методы исследования означают применение в большей или меньшей мере вычислительных технологий.

В силу ограничений, накладываемых диахроническим аспектом исследований, в качестве основных материалов для диахронической лексикографии выступают словари и различные первичные письменные тексты. Возможность произвольного сочетания корпусно-контекстологических и структурных методов со словарями или первичными текстами в качестве материала доказана работами многих авторов. Однако соображения производительности и эффективности диктуют нежелательность попыток вскрытия внутренней семантической структуры слова по данным корпуса текстов. Попыты автоматического извлечения данных из словарей пока что не дали возможности использования сложных семантических данных, содержащихся в них.

В настоящей работе анализ словарных определений в генетически связанных словарях разных лет применяется для обнаружения изменений в содержании одной единицы плана выражения, а корпусная методика обеспечивает обнаружения диахронических расхождений в плане выражения одной единицы плана содержания (поиск диахронических синонимов).

Сравнение словарей различных эпох может дать материал для выяснения закономерностей развития состава лексики языка как в количественном, так и в качественном отношении.

Современная лексикография располагает большим числом подробных классификаций типов и функций словарей. В целях определения круга словарей, которые могут быть привлечены в качестве материала для диахронических исследований лексической семантики, рассматриваются следующих типов: энциклопедические и толковые словари, идеологические словари, тезаурусы, нормативные и терминологические словари, учебные словари, переводные словари.

Хотя разделение энциклопедической и толковой функции в словаре редко проводится последовательно, энциклопедические словари, ориентированные на отражение научной, а не наивной картины мира, содержат меньше собственно языковой информации.

Нормативный характер тезаурусов и идеографических словарей ограничивает их применимость для диахронической лексикографии. Русскоязычные тезаурусы могут быть использованы для диахронического анализа глубиной не более 40 – 30 лет.

Применимость переводных словарей для изучения эволюции лексической семантики одного языка была продемонстрирована М.В.Марчук [Марчук, 1996]. Этот подход позволяет исследовать сдвиги на определенном фрагменте лексической системы, претерпевшем существенные изменения в одном языке, но оставшемся относительно стабильным в другом языке. Основой сравнения являются переводные эквиваленты, занимающие промежуточное положение между лексико-семантическим вариантом («значением») слова и компонентом значения (семой). Учитывается как появление и исчезновение переводных эквивалентов за исследуемый период, так и изменение места каждого в словарной статье. Описанный метод и основанная на нем теория диахронического вектора слова обладают высокой степенью формализации.

По сравнению с этой методикой, использование толковых словарей вносит определенную специфику и открывает новые возможности. При использовании в качестве элементарной единицы сравнения словарного значения, из-за существенно меньшего числа значений в толковом словаре, чем переводных соответствий в словаре двуязычном, а также из-за сложности их сопоставления, снижается надежность анализа. Использование же в качестве основы сравнения слов и синтаксических структур, составляющих толкование, открывает возможность извлечения семантической информации, недоступной в словарях переводных, хотя и повышает трудоемкость исследования.

Словарные статьи могут нести в себе информацию о всех трех уровнях семантики слова: сигнификативном, денотативном и коннотативном. Традиционное для толковых словарей определение через ближайший род и видовое отличие позволяет извлечь формулировку, соответствующую сигнификату, из большинства статей. В работе И.В. Рябовой [Рябова, 1996] в этих целях используются слова-понятия в субъектно-предикатной структуре определения.

В той же работе в качестве поверхностной реализации предметного содержания словарного определения рассматриваются слова-обозначения реальных объектов, явлений, признаков, действий, лиц и т.д. Однако референт совпадает с денотатом только при актуализации значения слова в речи. Кроме редких случаев, извлечение денотативного значения слова из словарного определения невозможно без привлечения концептуальной картины мира исследователя, так как именно она отделяет набор реально существующих объектов, включаемых в сигнификат, от потенциального объема денотата.

Исходя из определения коннотации, данного Ю.Д. Апресяном [Апресян, 1974], к признакам коннотационных компонентов определения, предлагаемых И.В. Рябовой (наречия, прилагательные, существительные оценочной семантики, слова-усилители типа «очень», «самый» и т.д.) добавляется следующее правило, специфичное для диахронического исследования: при относительной сохранности сигнификативного значения энциклопедическая информация, связанная со словом в более ранний период, допускается его толкованием и в более поздний; если какая-то дополнительная (не сигнификативная или денотативная) информация словарного определения стала неприменимой к более позднему толкованию, констатируется наличие коннотацией.

Исходя из представления о структуре определения в толковом словаре как о графе или дереве зависимостей, а не линейной иерархической структуре, предлагается использовать при сопоставлении словарных определений разных

лет функциональный подход, и выделяются следующие типы возможных изменений семантики слова:

Расширение сигнификата – в случае, когда определение конечного периода сместилось вверх по гипо-гиперонимическому ряду, либо если оно позволяет включить некоторые объекты в дополнение к исходному периоду. В любом случае понятие конечного периода полностью включает в себя понятие исходного периода.

Сужение сигнификата – аналогично, но при полном включении понятия конечного периода в понятие исходного периода.

Модификация сигнификата – любые другие изменения.

Аналогично определяются расширение, сужение и модификация денотата. С точки зрения коннотата предполагаются пейорация, элевация и иная модификация коннотата.

Изменение значений слов при неизменности означающих поддается регистрации легче остальных видов номинативных сдвигов. Однако установление различных по форме слов, которые могут быть признаны синонимами на определенном уровне своего значения, представляет особый интерес для исторического исследования.

В отсутствие готовых исторических словарей поиск «диахронических синонимов» неизбежно опирается на тексты соответствующих периодов. Так, в этих целях могут использоваться различные списки одного текста или переводы древнего текста на современный язык. Применение к таким текстам методов статистики и линейного программирования позволяет найти новые решения некоторых классических задач прикладной лингвистики, радикально сокращающих трудоемкость работ.

Первые работы с параллельными текстами были выполнены в конце 80-х – начале 90-х годов в рамках разработки различных систем *статистического машинного перевода*. [Harris, 1988], [Brown et al, 1990]. Первые методики машинного перевода на основе примеров послужили

толчком для дальнейших исследований в области извлечения лингвистической информации из параллельных текстов.

Именно в области автоматического составления словарей исследования по статистической обработке параллельных текстов достигли наиболее наглядных успехов. Так, некоторые системы достигают точности установления переводных соответствий на уровне слов выше 95% [Langlais et al, 1998].

Для обработки параллельных текстов исследователю необходимо решить несколько задач, а именно: подбор материала, т.е. собственно корпуса параллельных текстов, установление соответствия каждого предложения одного текста предложению, нескольким предложениям или части предложений другого текста (выравнивание предложений), поиск лексических соответствий, возможно, собственно статистический машинный перевод, и наконец, оценка результатов.

Применительно к экономической лексике, возможно использование ряда экономических текстов, написанных на иностранном языке и переведенных на русский язык в разное время. Существует не менее шести независимых переводов «Исследований о природе и причинах богатства народов» А. Смита, не менее четырех переводов собрания сочинений Д. Рикардо. Рассматривая переводы разных лет как параллельные тексты на разных языках, с помощью статистических методов можно составить словарь соответствий между переводами, относящимися к разным периодам.

Подавляющее большинство существующих параллельных текстов не дают возможности непосредственного извлечения информации о переводных соответствиях. Во-первых, очевидно, что между текстом оригинала и переводом не существует однозначного и линейного соответствия на уровне слов. Во-вторых, переводчик не всегда выбирает наиболее буквальный вариант перевода. Наконец, неизбежным фактом являются неточности перевода. Таким образом, любая работа по статистическому анализу параллельных текстов опирается на *выравнивание параллельных текстов*.

Наиболее популярный алгоритм выравнивания известен как алгоритм Гейла. Авторы алгоритма исходят из того, что длина предложений текста на языке перевода находится в прямой зависимости от длины предложения на языке оригинала [Gale, Church, 1991], а поиск оптимального варианта выравнивания выполняется путем *динамического программирования*, которое обеспечивает нахождение минимального суммарного расстояния между элементами двух последовательностей.

Альтернативный подход описан в работе [Chen, 1993]: в процессе выравнивания предложений алгоритм строит простую статистическую модель дословного перевода и выбирает такую схему выравнивания предложений, которая максимизирует вероятность генерации корпуса с этой моделью перевода, а функция расстояния вычисляется на основе произведения вероятностей лексических соответствий в рассматриваемой паре предложений. Таблица-словарь, а также остальные параметры модели перевода, вычисляются постепенно с помощью особой модификации алгоритма Витерби, при которой параметры модели пересчитываются не на всем корпусе, а в течение одного прохода по мере продвижения.

Параллельные тексты на одном языке дают возможность использовать промежуточный подход к выравниванию предложений. При расстоянии в несколько десятков лет между переводами одного текста сохраняется значительное число слов, не изменивших свое написание и значение, или изменившихся несущественно. Вследствие этого набор совпадающих или близких слов в предложениях двух текстов может служить показателем близости этих предложений, причем его уровень надежности существенно превышает надежность такого параметра как близость длины предложений.

Таким образом, можно построить алгоритм выравнивания, аналогичный алгоритму Гейла, т.е. использующий динамическое программирование без ограничения массива поиска, в котором в качестве функции расстояния используется информация о лексических соответствиях, но который является не рекурсивным (как алгоритм Витерби), а детерминистским.

При отсутствии дополнительных лингвистических ресурсов наиболее доступным способом лемматизации является извлечение слов из рассматриваемых текстов в тех формах, в которых они использованы, с последующей оценкой вероятности того, что две словоформы являются одним словом на основании отношения совпадающих последовательностей знаков и различающихся. Такой подход оправдан, в частности, при сопоставлении текста в современной и устаревшей орфографии.

Два параллельных корпуса, выровненные на уровне предложений, позволяют извлечь информацию о переводных соответствиях двух языков, что выполняется путем построения лексических конкордансов. Наиболее простым статистическим алгоритмом построения лексических конкордансов является метод, предложенный И. Меламедом [Melamed, 1996]. Однако недостаток этого метода состоит в том, что степень связанности пар слов вычисляется независимо друг от друга. Это приводит к появлению в конкордансе не только пар, соединенных непосредственной связью, но и пар с косвенной связью.

Более точные результаты достигаются путем построения *модели перевода*, когда за основу берется модель, позволяющая оценить правдоподобие конкретной гипотезы о схеме переводных соответствий в данных параллельных текстах, а затем методами линейного программирования ищется такая схема соответствий, которая обладает максимальной оценкой.

При статистическом подходе проблема перевода рассматривается в терминах *канала с помехами*. Согласно принципу канала с помехами предложение на одном языке представляет собой предложение на другом языке, искаженное неким шумом. Перевод осуществляется путем поиска такого предложения, которое максимизирует произведения безусловной вероятности предложения перевода и вероятности предложения оригинала при условии данного предложения перевода. Для статистического машинного перевода требуется *модель источника* и *модель канала*, или *модель языка* и *модель перевода*.

В качестве *модели языка* в системах статистического перевода используются преимущественно различные модификации n-граммной модели, утверждающей, что грамматичность выбора очередного слова при формировании текста определяется только тем, какие $(n - 1)$ слов идут перед ним. Вероятность каждого n-грамма определяется по его встречаемости в тренировочном корпусе.

Самой простой статистической *моделью перевода* является модель дословного перевода. В этой модели, известной как Модель IBM №1, предполагается, что для перевода предложения с одного языка на другой достаточно перевести все слова (создать «мешок слов»), а расстановку их в правильном порядке обеспечит модель языка. Обучение Модели №1 производится на корпусе параллельных текстов, выровненном на уровне предложений. Описание математической части Модели №1 подробно изложено в работе [Knight, 1999].

Слабость Модели №1 демонстрируется на простом примере: представим, что модель перевода предоставила следующий «мешок» слов: утром, вечером, деньги, стулья. С точки зрения модели языка предложения «Утром деньги, вечером стулья» и «Утром стулья, вечером деньги» будут, по-видимому, иметь примерно равную вероятность, однако один из вариантов явно искажает смысл.

Более поздние модели статистического перевода (№2 – №5) решают ряд задач по передаче синтаксической конструкции переводимого предложения. В модели №4 появляется понятие класса слов. Более совершенные модели обеспечивают более высокую точность составления лексического конкорданса, однако это улучшение не столь значительно, чтобы оправдать использование существенно более ресурсоемких методов для решения задачи извлечения лексических соответствий.

Наиболее распространенными критериями оценки результатов статистической обработки текста являются *точность* (precision) и *охват* (recall). Измерение этих величин предполагает сравнение результата работы

оцениваемой системы с эталоном. Для этого выбирается единица подсчета (например, предложение), и все единицы оцениваемого результата сравниваются с эталоном. Точность определяется как отношение числа правильных единиц результата к общему числу единиц в оцениваемом результате. Охват определяется как отношение числа правильных единиц результата к общему числу единиц эталона.

Для проверки практической применимости положений, изложенных в первой главе, были проведены два эксперимента, позволившие не только отработать новые приемы диахронического исследования экономической лексики русского языка и построить набор обеспечивающих их инструментов, но также получить определенный массив собственно диахронического лексикографического материала.

В рамках первого эксперимента была осуществлена попытка изучения эволюции семантики сплошной выборки экономической лексики по данным двух толковых словарей русского языка. Содержанием второго эксперимента стал поиск диахронических синонимов в двух переводах одного иноязычного текста с помощью методов, основанных на современных работах по статистическому машинному переводу.

Для анализа сдвигов в семантической структуре экономической лексики русского языка в 20 веке в качестве словарей-источников были выбраны словари толкового типа. Ближайшим к началу исследуемого 20 века и подходящим по своим параметрам словарем оказался «Толковый словарь русского языка», составленный под редакцией Д.Н. Ушакова в 1927 – 1940 годах [Толковый словарь ..., 1935-1940], соединивший в себе традиции строгой научности академических словарей и с общедоступность массового словаря-справочника, предполагающую строго нормативный характер словаря.

Со словарем Д.Н. Ушакова генетически тесно связан словарь С.И. Ожегова, работа над которым началась в 1940 году, а первое издание которого вышло в свет 1949 году. Несмотря на то, что объем словаря С.И.

Ожегова составляет один том против четырех томов словаря Ушакова, даже в его первых, менее объемных изданиях размер словника (52 – 53 тысячи слов) составляет почти две трети словника словаря Ушакова (около 85 тысяч слов).

Вышедшее в свет в 1991 г. 23 издание словаря С.И. Ожегова с объемом словника около 70 тысяч единиц было выбрано в качестве источника, представляющего ситуацию в нормативном русском литературном языке на конец 20 века [Ожегов, 1991].

Для отнесения лексики, представленной в выбранных словарях, к сфере экономики использовался Тезаурус по экономике и демографии ИНИОН [Архангельская, 2001]. Из 314 слов, присутствующих и словаре С.И. Ожегова, и в тезаурусе, были отобраны 210 слов, относящихся только к сфере экономики. При этом из списка были исключены слова, относящиеся к демографии (напр., «женщины», «заболеваемость»), к метаязыку науки (напр., «динамика»), а также такие слова, которые не имеют лексико-семантических вариантов (ЛСВ) с экономическим значением ни в одном из словарей (напр., «лаг», «мультипликатор»).

В соответствии с положениями, изложенными в первой части работы, при сопоставлении словарных статей использовались следующие принципы:

- В большинстве случаев одно словарное значение трактовалось как один ЛСВ. В отдельных случаях имеют место полноценные самостоятельные определения, включенные в состав других определений; такие определения рассматриваются как отдельные ЛСВ. В случае отсылочного толкования использовалось толкование того слова и значения, на которое указывает отсылка.
- Отнесение отдельных ЛСВ к сфере экономики производилось экспертным способом. Экономические ЛСВ одного заглавного слова в каждом из двух словарей соотносились друг с другом на основании экспертной оценки близости определений. Анализ ЛСВ, не признанных экономическими, в большинстве случаев не производился. Соотнесение неэкономических ЛСВ

осуществлялось только если это было необходимо для определения актуализации/деактуализации экономических значений.

- Занятие отдельным толкованием более высокой (или низкой) позиции среди других определений интерпретировалось как актуализация/деактуализация соответствующего ЛСВ. Учитывался не абсолютный номер словарного толкования, а его положение относительно других ЛСВ – в первую очередь экономических.
- Извлечение из словарной дефиниции сигнификативного компонента значения производилось путем анализа его субъектно-предикатной структуры и ограничивающих определений.
- Сравнение денотативного значения слова по данным двух словарей производилось путем интуитивной оценки возможности включения в класс определяемых объектов различного набора таковых с обеих сторон диахронической пары.
- Если часть определения признавалась энциклопедической информацией, она исключалась из рассмотрения.
- Если по каким-либо причинам можно было установить большую или меньшую экономическую соотнесенность понятия в одном из периодов, такое изменение отмечалось отдельно.

Сравнительная схема определений словаря Д.Н. Ушакова 1935-1940 г.г. и словаря С.И. Ожегова 1991 г. приводится в Приложении 1. В левой колонке приводится статья словаря Ушакова, в правой – словаря С.И. Ожегова. Справа от заглавного слова, а также слева от каждого толкования могут приводиться условные обозначения отмеченных изменений семантики; отсутствие таких помет означает полное совпадение семантики по толкованиям обоих словарей.

Проведенный эксперимент позволяет установить возможность исследования эволюции семантики основного терминологического слоя экономики с помощью толковых словарей русского языка разных лет. Этот подход эффективен для изучения истории означающих тех слов, которые в настоящее время входят в сферу экономической лексики. Он позволяет

установить достаточно тонкие качественные сдвиги в лексической семантике на большинстве ее уровней.

Анализ генетически близких и схожих по лексикографическим параметрам словарей, разделенных периодом около 50 лет показывает, что даже за такой относительно небольшой отрезок времени в той части литературного русского языка, которая пересекается с терминологическим полем экономики, происходят многочисленные сдвиги.

Так, из 210 слов, зафиксированных в словаре Ожегова, 13 являются новыми, у 24 слов появились новые экономические ЛСВ. С другой стороны, из зарегистрированных экономических ЛСВ в словаре Ушакова 42 отсутствуют в словаре Ожегова, что связано как изменениями на семантическом уровне русского языка, так и с различием объема двух словарей.

У 62 экономических ЛСВ зарегистрированы сдвиги на уровне сигнификата, из них 30 – в сторону расширения, у 52 экономических ЛСВ зарегистрированы сдвиги на уровне денотата, из них 28 – в сторону расширения. Тогда как не все расширения денотата/сигнификата могут считаться достоверными в силу соотношения объема словарей, количество относительно достоверных сдвигов остается весьма существенным. У экономических 17 ЛСВ зарегистрированы изменения коннотата, у 11 экономических ЛСВ – изменения в уровне актуальности их экономического значения, и наконец, у 29 экономических ЛСВ – изменения их актуальности в семантической структуре слова.

Для изучения возможности применения методов статистической обработки параллельных текстов к диахроническому анализу русской экономической лексики была выбрана работа А. Смита «Исследование о природе и причинах богатства народов». В качестве параллельных текстов были выбраны два перевода – 1895 года и 1931 года.

Текст перевода 1895 года доступен только в печатном виде. Для машинной обработки текст было необходимо отсканировать и произвести его

распознавание. Из-за отсутствия поддержки дореволюционной орфографии в существующих программах оптического распознавания текста количество ошибок при распознавании оказалось очень высоким – от единиц до десятков на страницу. Исправление этих ошибок было произведено вручную. В процессе исправления также были удалены многочисленные комментарии, не входящие в основной текст. Таким образом перевод 1895 года был преобразован в машиночитаемый текст объемом 325 571 знак.

Текст перевода 1931 года доступен в электронном виде, что существенно облегчает работу с ним. Исходный объем этого текста составляет 535 665 знаков.

Для дальнейшей работы с текстами использовался набор программных инструментов, разработанных автором на платформе Microsoft Access и Microsoft Visual Basic и выполняющих следующие функции, необходимые для поиска лексических соответствий в двух параллельных массивах текста на русском языке:

- Разделение текста на предложения
- Разделение предложений на слова
- Выравнивание предложений двух массивов между собой

Для разделения текста на предложения был применен простой алгоритм, использующий в качестве признака конца предложения знаки точки, вопросительный или восклицательный знак, после которых следует пробел или конец абзаца. Для разделения предложений на слова был применен алгоритм, использующий в качестве признака конца слова любой знак, не являющийся буквой русского алфавита. Для выравнивания предложений использовался алгоритм динамического программирования без ограничения области поиска. Функция расстояния использовала сочетание числа лексических совпадений и близости длин предложений.

Для нахождения лексических совпадений предварительно был построен список словоформ, считающихся формами одного слова. Для этого все словоформы, найденные в корпусе 1895 года и в корпусе 1931 года, были

сравнены с точки зрения их графической близости. Формами одного слова считались те кандидаты, коэффициент схожести которых превысил 0,8. Для того, чтобы снизить вероятность ситуации, когда два предложения, являющихся переводом одного предложения оригинала, было невозможно сопоставить, потому что в них не было обнаружено графически достаточно близких словоформ, в дополнение к мере лексического сходства использовалась мера сходства длины, заимствованная из исходного алгоритма Гейла. В реализации алгоритма динамического программирования Гейла для операций удаления, одностороннего слияния и двустороннего слияния предложений использовались «штрафы», соответствующие вероятности операции каждого типа в корпусе текста.

При просмотре полученной таблицы было обнаружено, что текст 1895 года перед обширными пропусками имеет характер скорее реферата, чем перевода. Для снижения уровня шума на следующем этапе эксперимента единичные предложения, сопоставленные со случайными предложениями на фоне обширных пропусков, были удалены вручную. В результате были получены 1.076 пар предложений.

Для эксперимента была выбрана Модель IBM №1. В ходе эксперимента был использован программный код, в работе [Statistical Machine Translation, 1999]. Результатом эксперимента стала таблица лексических соответствий, расположенных в порядке убывания вероятности, присвоенной паре статистической моделью. Конечный этап работы заключался в ручном просмотре пар-кандидатов с вероятностью выше 0,5 и коэффициентом графической схожести ниже 0,8. Из 295 пар были отобраны пары слов, являющихся синонимами, после чего автором были вручную отмечены слова, относящиеся к области экономики.

По результатам проведенного исследования были сделаны следующие выводы:

1. Лексическая система характеризуется полисемией, к исследованию которой возможны контекстологические, структурные, психолингвистические

и корпусные подходы. В настоящее время наиболее активно развиваются методы, основанные на обработке обширных корпусов текстов и сочетающие в себе элементы структурного, контекстологического и психолингвистического подходов.

2. Методы и инструменты исследования лексической полисемии могут рассматриваться как потенциальный источник средств исследования исторических лексико-сематических изменений.

3. Рассмотрение сдвигов отношений номинации необходимо проводить не только относительно единиц плана выражения, но и в ракурсе вариативности плана выражения относительно констант плана содержания.

4. Толковые словари представляют собой один из наиболее эффективных инструментов анализа эволюции семантики русской лексики в 20 веке. Данные толковых словарей разных лет позволяют установить такие факты как а) появление, развитие и исчезновение лексической единицы в языке; б) приобретение и утрата словом отдельных значений (ЛСВ), а также вхождение слова в определенный подъязык и выход из него; в) приобретение и утрата отдельным лексико-семантическим вариантом компонентов и оттенков значения на сигнификативном, денотативном и коннотативном уровне.

5. Генетически связанные словари Д.Н. Ушакова 1935-1940 г.г. и словаря С.И. Ожегова редакции 1991 года позволяют констатировать различные изменения в семантике более трети зарегистрированных экономических лексем за соответствующий период.

6. Для поиска диахронических синонимов по переводам одного иноязычного текста, выполненным в разные годы, может использоваться методика, основанная на модели статистического перевода IBM №1.

7. Для распознавания печатного русскоязычного текста в дореволюционной орфографии с существенными ограничениями может использоваться программный продукт «ABBYY FineReader».

8. Для выравнивания предложений двух переводов иноязычного текста в качестве меры схожести предложений эффективно использование обобщенного показателя количества лексических совпадений и близости длин.

9. Для приведения словоформ с потенциальными отличиями орфографии и возможными опечатками к лемме возможно использование коэффициента схожести на основе числа совпадающих и различающихся букв.

10. Изменение общепринятого наименования ряда экономических реалий за первую треть 20 века привело к замене автором перевода «Исследований» А.Смита 1931 года не менее 65 экономических лексем на варианты, отличные от первого перевода 1895 года.

Основные положения диссертации отражены в следующих публикациях автора:

1. *Рахимбердиев Б.Н.* Об эволюции семантики некоторых экономических терминов. // Сборник трудов X сессии Российского Акустического Общества – М.: Российское Акустическое Общество, 2000 – 0,15 п.л.

2. *Рахимбердиев Б.Н.* Применение статистических методов в компьютерной лексикографии. // Русский язык: исторические судьбы и современность. Труды и материалы международного конгресса – М.: Изд-во МГУ, 2001 – 0,1 п.л.

3. *Рахимбердиев Б.Н.* Статистическая обработка параллельных текстов. // Акустика речи и прикладная лингвистика. – М.: Изд-во МГЛУ, 2002 – Вып. 3 – 1,3 п.л.