УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 152, кн. 2 Гуманитарные науки *2010*

УДК 811.161

МОРФОСИНТАКСИЧЕСКИЙ МЕТОД ИЗВЛЕЧЕНИЯ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИИ ИЗ КОРПУСОВ

В.Д. Соловьев, В.Р. Байрашева

Аннотация

Статья посвящена описанию нового метода извлечения информации о семантике слов из корпусов текстов. Метод основан на выделении синтаксических отношений, в которые входит изучаемое слово в текстах, описании морфологической маркировки этих отношений и получении частотных характеристик маркировок. Метод ориентирован на то, чтобы извлекать информацию из текстов, опираясь на формальные признаки, которые просто и безошибочно различаются, и избегать привлечения семантической интуиции исследователя. Возможности метода демонстрируются на примере проблемы синонимии. Обобщаются результаты, полученные ранее в ряде наших работ, приводятся новые данные. Другим возможным применением метода является описание метафор и структуры семантических пространств.

Ключевые слова: корпус, семантика, статистический анализ, синонимия, лексика семантического поля *эмоции*.

Введение

С созданием больших корпусов текстов появилась возможность по-новому взглянуть на многие языковые проблемы, в частности на проблемы семантики. Вопросам корпусного исследования семантики и других уровней языка посвящены вышедшие в последние годы работы [1, 2]. Корпусный анализ может оказаться весьма эффективным методом и при описании синонимов.

Несмотря на наличие значительного числа работ по синонимии, проблема точного описания синонимов еще далека от разрешения. Существующие определения синонимии не являются строго операционными, то есть не позволяют по двум словам автоматически определить, являются ли они синонимами. Одна из целей данной работы – предложить возможный вариант такого определения. О трудности описания синонимов свидетельствуют большие расхождения в словарях синонимов у различных авторов. Это означает, что интуиции исследователей недостаточно для получения общепризнанного решения по данной проблеме и должны быть предложены новые объективные методы определения и анализа языковых синонимов. Изучение синонимических рядов представляется интересным в теоретическом плане, так как оно дает информацию о способах организации семантического пространства языковым аппаратом человека.

Наше исследование базируется на следующей общей идее: семантика определяет синтаксис. Хотя данный тезис в какой-то степени противоречит принципиальным теоретическим установкам Хомского о независимости синтаксического модуля, он поддерживается некоторыми исследователями, в частности

А. Вежбицкой. Если общая идея верна, это означает, что близкие по смыслу слова – синонимы – должны обладать и близкими синтаксическими свойствами. В итоге реализации предложенного подхода должно получиться следующее определение синонимии:

(#) Два слова (одного семантического поля) являются синонимами, если они обладают близкими синтаксическими свойствами.

Сравним его с традиционным определением синонимов как слов, близких по смыслу. Оба определения не являются строгими, так как не уточняют, что означает «близкие». Однако определение (#) является более строгим и легче верифицируемым за счет того, что оно апеллирует к синтаксису, а не к значению слова. Синтаксис же является значительно более четко структурированным уровнем языковой системы, чем семантика. Он достаточно точно и полно описан (по крайней мере для основных языков). В отношении синтаксической правильности предложений обычно не возникает разногласий у носителей языка. Кроме того, синтаксические связи обычно маркируются на поверхностном уровне легко различимыми средствами – падежами, предлогами, порядком слов. Все это позволяет достаточно легко и с высокой степенью достоверности извлекать синтаксическую информацию из текстов. В схожем по идеологии корпусном исследовании семантики [3] также делается упор на извлечении из текстов легко вычленяемой информации, преимущественно синтаксической и морфологической. Хотя в определении (#) используется понятие семантического поля, это не требует проведения тщательного семантического анализа: для перехода к рассмотрению синтаксических свойств слов достаточно приближённого выделения групп лексем с общим значением (например, «эмоции»).

Другое часто используемое определение синонимии основано на идее взаимозаменяемости синонимов, по крайней мере в части контекстов. При таком подходе остается неясным, что считать контекстом. Наше определение фактически уточняет также и понятие контекста, трактуя его как набор тех слов в предложении, с которыми исследуемое слово находится в синтаксических отношениях.

Первая проблема состоит в том, чтобы точно специфицировать, какую же именно синтаксическую информацию надо извлекать из текстов. Затем идет стадия обработки текстов и, наконец, стадия статистической обработки результатов.

Один из возможных подходов к корпусному анализу семантики реализован в LSA – Latent Semantic Analysis [4]. В LSA с каждым вхождением слова в тексте связывается его ближайшее окружение, то есть набор близко расположенных слов. Затем по всем вхождениям данного слова в корпус текстов подсчитывается, сколько раз появляются различные слова в его окружении. На основании полученных результатов строится матрица корреляции между словами и контекстами, в которых они появляются. В дальнейшем матрица подвергается обработке с помощью математического алгоритма декомпозиции, близкого к используемым в факторном анализе. Можно считать, что в LSA смысл слова представляется как средний смысл всех контекстов, в которых оно встречается.

Программы, работающие на основе LSA, применимы и к проблеме синонимии. Как отмечено в [4], при определении синонимов они демонстрируют результаты, отражающие языковую компетенцию студентов американских вузов, для которых английский язык не является родным.

Особенностью, вероятно, недостатком, LSA является то, что в нем совершенно не учитывается синтаксис – ни порядок слов, ни морфологическая маркировка. Одна из причин этого состоит в сложности автоматической обработки предложений с учетом синтаксиса, другая – в бедности английской морфологии. Однако, как признают и сторонники этого подхода [4], игнорирование синтаксиса приводит к ошибкам. Так, в [4] приведен забавный пример, когда в результате LSA близкими по смыслу совершенно необъяснимо признаны слова *sadomasochism* ‘садомазохизм’ и *verbally* ‘словесно’!

1. Объект исследований

Метод апробировался на нескольких группах слов русского языка, в том числе на группе лексем, обозначающих эмоции (см. табл. 1).

Выбор объекта исследования обусловлен следующими факторами. Прежде всего русский язык выбран потому, что существуют обширные свободно доступные корпуса текстов: Национальный корпус русского языка, сокращенно НКРЯ (http://ruscorpora.ru), и библиотека Машкова (http://aot.ru), содержащие более 100 млн. слов и более 600 млн. слов соответственно. Правда, библиотека Машкова включает только художественную литературу и по этому признаку является менее сбалансированной, чем НКРЯ. Однако наличие двух больших корпусов позволяет повысить степень достоверности результатов за счет возможности сопоставления полученных из них данных.

Кроме того, русский язык отличается богатой морфологией, аналогичное исследование на материале английского языка было бы затруднено. Еще одна из причин выбора русского языка состоит в том, что упомянутое выше исследование [3] также проводилось на материале русского языка, что дает возможность напрямую сопоставить данные методы. Выбор существительных позволяет расширить методику по сравнению с подходом Д. Дивьяка, примененным к глаголам. Эти работы являются в какой-то мере дополняющими друг друга.

Выбор слов, обозначающих эмоции, вызван тем, что это чрезвычайно сложная область, пожалуй, максимально трудная для описания синонимии. В ней определенно нет точных синонимов, то есть слов, имеющих один и тот же денотат. Слова из этой области, считающиеся синонимами, обозначают близкие, но различные эмоциональные состояния. Кроме того, эта область давно изучается и имеется возможность сопоставления результатов, полученных нашим методом, с результатами, полученными с помощью обычной интроспекции исследователя – лексикографа или специалиста по семантике. Наконец, выбор слов осуществлен так, что большая часть из них более или менее соответствует областям, покрываемым в английском языке словами *sadness* ‘печаль’ и *fear* ‘страх’ и тщательно изученным в англоязычной литературе, что дает широкие возможности для проведения сопоставительных исследований.

Для каждого исследуемого слова с помощью поисковых систем вышеназванных корпусов выбирались содержащие это слово предложения. Для каждого слова было взято и проанализировано не менее 500 предложений (первых, выданных поисковой системой соответствующего корпуса). Анализ предложений осуществлялся студентами-филологами 3–4 курсов Казанского государственного университета и состоял в выделении синтаксических отношений.

Табл. 1

Частота основных маркировок дополнений и обстоятельств (тип 2 синтаксических отношений)

Лексема	пред + *в*	вин + *в*	твор	твор + *с*	род + *от*	+	вин
страх	10	1	5	10	20	21	35
испуг	19	2	14	22	22	12	10
трепет	8	7	4	50	1	6	22
ужас	21	14	8	19	11	17	12
волнение	8	1	4	17	35	10	25
растерянность	40	5	4	3	10	10	28
скорбь	9	3	5	12	3	26	44
грусть	2	2	9	55	1	13	17
печаль	7	5	10	16	5	17	41
тоска	6	3	12	25	14	12	30
разочарование	1	1	14	6	6	27	46
депрессия	6	42	11	2	8	19	12
отчаяние	20	17	2	12	14	20	16
уныние	11	41	5	5	4	8	26
меланхолия	7	23	20	2	9	16	23
хандра	7	21	7	14	21	14	14
горе	10	6	11	5	10	27	26

2. Проблема определения синонимии

Рассмотрим проблему синонимии на примере набора {*грусть*, *печаль*, *тоска*, *уныние*, *меланхолия*, *хандра*}. В различных словарях синонимов и толковых словарях это множество слов структурируется по-разному. Фактически общепринятым является объединение слов *грусть*, *печаль*, *тоска* в один синонимический ряд. Выделяется еще один синонимический ряд – *меланхолия*, *хандра*. Очень сложной является ситуация со словом *уныние.* В [5] оно считается синонимом *печали*, в [6] – синонимом *хандры*, а в [7] – помещено сразу в оба ряда: и в ряд со словом *грусть,* и в ряд со словом *хандра*, то есть фактически постулируются два значения этого слова. Последнее решение нельзя признать вполне удачным, так как очень трудно выделить и дать толкования двум разным значениям слова *уныние*. Например, в толковом словаре [8] указывается только одно значение этого слова.

При построении толкований различные авторы обращают внимание на различные аспекты эмоций. Например, в [5] для всего ряда {*грусть*, *печаль*, *тоска*} дается следующее толкование ‘неприятное чувство, какое бывает, когда нет того, что человек хочет, и когда он думает, что желаемое невозможно’. Таким образом, основу дефиниции здесь составляет указание причины возникновения такого рода эмоций[1]. Альтернативный подход можно найти в [8], где печаль определяется как ‘скорбно-озабоченное, нерадостное, невеселое настроение, чувство’. Здесь упор делается на самом чувстве. Систематическое применение этого подхода наталкивается на ту трудность, что мы просто не располагаем необходимым метаязыком для описания чувств как таковых. Фактически здесь приходится

[1] Классификация эмоций по причинам, их вызывающим, характерна и для ряда психологических работ [9].

описывать одни эмоции через другие. Рядом авторов [10, 11] предпринимались попытки использования метафоры как средства толкования лексем, называющих эмоции. Так, Ю.Д. Апресян предлагал в словарную статью для слова *страх* ввести такое толкование: ‘душа человека чувствует нечто подобное тому, что ощущает его тело, когда ему холодно; тело реагирует на это как на холод’. К сожалению, хорошие метафоры этого типа имеются лишь для небольшого числа эмоций и развить эту метафору до полномасштабного языка толкования эмоций не удалось.

В большинстве словарей синонимов приводится лишь описание целых рядов слов, но нет строгой дифференциации их значений внутри ряда. Исключением является словарь [5], но он, к сожалению, содержит весьма ограниченное число слов. Не вполне ясным является и набор признаков, по которым следует дифференцировать эмоции.

В идеале для получения строгого определения синонимии в формулировке (#) требуется уточнить, что значит «близкие». Для этого необходимо получить некоторую количественную меру близости лексем. Традиционные словари не содержат никаких количественных характеристик слов. Получение надежных количественных характеристик стало возможным лишь недавно, после появления больших корпусов текстов. Одним из основных результатов цикла наших исследований является определение перечня параметров, описывающих синтаксическое поведение лексемы, и нахождение с использованием корпусов текстов количественных характеристик этих параметров, которые могли бы быть учтены в толкованиях слов. Количественные характеристики могут быть также использованы для определения степени близости слов.

3. Морфосинтаксические параметры

Семантико-синтаксические свойства лексемы проявляются в различных формах – в форме синтаксических связей, в которые способна вступать данная лексема, в форме участия лексемы в различных синтаксических процессах (например, пассив для глаголов) и в форме сочетаемости с другими лексемами. Интересная попытка описания семантики глаголов через синтаксические процессы была предпринята в работе [12].

В упомянутой работе акцент делается на количественном описании синтаксических связей лексем. Для глаголов синтаксические связи отражаются в виде модели управления – наборе актантов (участников ситуации, описываемой глаголом). Этот подход легко распространить на часть существительных, обозначающих ситуации и являющихся так называемыми отглагольными именами. Однако для предметных имен типа *стул* такой подход не применим напрямую. Развиваемый нами подход можно трактовать как попытку распространить актантный способ описания на все существительные.

Реализация данного подхода предполагает прежде всего составление перечня синтаксических отношений, в которые вступает изучаемое существительное. Придерживаясь установившейся традиции описательных грамматик (русского языка), выделим следующие основные группы синтаксических отношений.

1. Предикативная связь: существительное в роли подлежащего – глагол.

2. Подчинительная связь в рамках клаузы: глагол – существительное. Здесь существительное выступает в роли дополнения или обстоятельства.

3. Подчинительная связь в рамках именной группы:

а) согласованные определения: существительное – модификаторы (прилагательные, местоимения, числительные);

б) несогласованные определения, в которых рассматриваемое слово выступает в роли определяемого;

в) несогласованные определения, в которых рассматриваемое слово выступает в роли определяющего.

4. Сочинительная связь.

На данный момент проведены исследования четырех групп синтаксических отношений – 2, 3б, 3в, 4. В настоящей статье приводятся числовые показатели по группам 2 и 3б.

Интересно посмотреть, на какие синтаксические отношения эмотивной лексики было обращено внимание в предшествующих исследованиях. Возвратимся к работе Ю.Д. Апресяна [10]. Метафора *страх – это холод* обосновывается 15 примерами, из которых в трех слово *страх* занимает позицию подлежащего и в остальных 12 – позицию косвенного дополнения (маркированного предлогом *от* и родительным падежом).

В другой классической работе по эмоциям в русском языке [11] обосновывалось метафорическое представление об эмоциях как жидкости. В этой статье из 57 примеров, обосновывающих данное положение, в 19 эмотивное слово подчиняется некоторому слову в составе именной группы (тип 3в), в 18 является подлежащим (тип 1) и в 30 занимает позиции дополнений и обстоятельств (тип 2). Из них в 9 случаях эмотивное слово маркируется творительным падежом, в 7 – винительным, в 5 случаях – пред. пад. + *в*, в 4 – род. пад. + *от*, в 2 – род. пад. и по одному – пред. пад. + *на*, вин. пад. + *за*, дат. пад.

Это указывает на важность выделенных выше групп 1, 2, 3в синтаксических отношений для понимания семантики эмоций.

Рассмотрим подробнее указанные группы синтаксических отношений, начиная со 2-й. Как в традиционных описательных грамматиках, так и в более современных грамматических теориях, таких, как реляционная грамматика Перлмуттера – Постала, дополнения делятся на прямое, непрямое и косвенные. Это соответствует 1-й, 2-й и 3-й валентностям в формализации синтаксических отношений в русском языке у Ю.Д. Апресяна [13]. Для целей описания семантики представляется уместным рассматривать косвенные дополнения не как единое синтаксическое отношение, а ввести некоторую их классификацию.

Косвенные дополнения различаются как по семантическим ролям, так и по их поверхностному оформлению с помощью предлогов и падежей. Использовать семантические роли напрямую в рамках данного подхода нежелательно, так как это потребует привлечения языковой интуиции, мы же стремимся извлекать информацию из текстов, опираясь только на формальные признаки, которые просто и безошибочно различаются. Таким образом, остается обратиться к предложно-падежной маркировке. Это тем более естественно, что она служит именно целям кодирования семантики.

Разумеется, каждый предлог и падеж имеет несколько (и даже много) значений. Например, комбинация ‘родительный падеж + *от*’ маркирует значение причины (*Он остолбенел от страха*), членимого целого (*Он отщипнул лучину*

от полена) и другие. Однако сужение предметной области исследований (в данном случае выбор лексем с общим значением *эмоции*) сильно ограничивает многозначность кодировки. В итоге у каждой поверхностной маркировки остается небольшое число значений, релевантных выбранной предметной области, а иногда только одно.

Таким образом, мы выделяем среди косвенных дополнений столько синтаксических отношений, сколько существует различных поверхностных маркировок. Аналогично поступаем и с обстоятельствами. В традиционных грамматиках обстоятельства классифицируются на основе семантики: обстоятельства места, времени и т. д. Отказавшись от использования семантики, нам не остается ничего иного, как классифицировать обстоятельства по падежно-предложной маркировке. Учитывая, что разграничение косвенных дополнений и обстоятельств не является строго формальным и не может быть осуществлено (по крайней мере в русском языке) на основе поверхностных синтаксических признаков, принято решение не дифференцировать эти категории.

Косвенные дополнения и обстоятельства, имеющие одинаковую предложно-падежную маркировку, попадут в один класс и будут рассматриваться совместно. Это решение отражает установку на привлечение только явно выраженных на поверхностном уровне признаков.

4. Демонстрационный пример

Прежде чем углубиться в область эмоций, рассмотрим один очень простой демонстрационный пример из области физического мира – слово *стул* – и увидим, что именно кодируют дополнения и обстоятельства и как вообще выглядит предлагаемая методика в целом.

500 предложений взято из Библиотеки Машкова. Из них в 358 (более 70%) *стул* занимал позицию дополнения или обстоятельства. В общей сложности встретилось 20 различных предложно-падежных маркировок[1]. Статистика выглядит следующим образом: вин + *на* – 108 случаев, пред + *на* – 82, род + *с* – 70, вин (вместе с род при глаголах с отрицанием) – 64, дат + *к* – 8, вин + *за* – 4, твор + *за* – 4, твор + *под* – 4, род + *от* – 2, твор – 2, твор + *с* – 2, вин + *в* – 1, род + *из* – 1, род + *около* – 1, род + *из-под* – 1, род + *позади* – 1, пред + *о* – 1, твор + *вместе*+*с* – 1, твор + *над* – 1.

Только 4 маркировки встречаются часто: вин + *на* – 30% всех маркировок, пред + *на* – 23%, род + *с* – 20% , вин – 18%. Все остальные маркировки вместе взятые – 9%, причем каждая из них в отдельности – не более 3%.

То, что часто встречается винительный падеж – прямое дополнение, – не удивительно, вероятно, так будет и для всех (или почти всех) других существительных в языках номинативного строя. Рассмотрим остальные три частотные маркировки. Типичный контекст для ‘вин + *на*’ – ‘Сесть на стул’, для ‘пред + *на*’ – ‘Сидеть на стуле’, для ‘род + *с*’ – ‘Встать со стула’. Таким образом, очевидно, что в выделенных предложно-падежных сочетаниях отразились основные спо-

[1] Далее будем сокращенно обозначать родительный падеж – род, винительный падеж – вин, дательный падеж – дат, творительный падеж – твор, предложный падеж – пред. Сочетание падежа с предлогом обозначается знаком «плюс».

собы использования стула в жизни людей. Конечно, это не является неожиданным, однако приведенные данные указывают на то, что наиболее существенные семантико-прагматические свойства концептов прямо влияют на частоту проявления определенных формальных показателей в текстах. Частота же легко вычисляется по корпусу текстов и является естественным источником объективной информации о семантике.

Аналогичные результаты можно ожидать и в более сложном случае – в сфере эмоций. Наиболее частотные маркировки должны отражать наиболее существенные глубинные семантико-прагматические свойства соответствующих концептов.

5. Статистические данные

Приведем статистические данные по выбранному множеству слов. Для изучаемого семантического поля наиболее частотными оказались следующие маркировки дополнений и обстоятельств: вин, пред + *в*, вин + *в*, твор, твор + *с*, род + *от*. Все остальные маркировки (дающие меньшие проценты) объединены вместе в столбце ‘+’. Данные приведены в табл. 1. Несогласованные определения чаще всего имеют следующие маркировки: твор + *перед*, род + *от*, пред + *в*, дат + *по*, род (см. табл. 2). В столбце ‘+’ объединены все остальные. Определения, маркированные родительным падежом, имеют существенно различную и легко различимую семантику: они либо специфицируют субъекта эмоции (*страх девушки*), либо причину ее появления, а также иные факторы (*страх высоты*). Возможно, эти два вида следует разделить, что и отражено в табл. 2. Правда, это является определенным отступлением от декларированного принципа неучёта семантики. Оправданием может служить то, что эти два вида семантики очень легко различить, ошибки, то есть влияние субъективной оценки, практически исключены. Впрочем, эти два столбца можно и объединить.

Табл. 2

Частота основных маркировок несогласованных определений (тип 3б)

Лексема	твор + *перед*	род + *от*	пред + *в*	дат + *по*	пред + *о*	+	род (суб)	род (др.)
страх	50	1	2	0	0	19	11	19
испуг	15	10	20	0	0	35	20	0
трепет	40	0	7	2	2	14	9	27
ужас	8	0	0	0	4	0	84	4
волнение	6	10	5	0	2	22	33	21
растерянность	20	7	5	0	0	5	63	0
скорбь	3	1	13	10	14	7	32	20
грусть	0	0	17	28	6	6	11	33
печаль	0	2	0	6	9	11	21	51
тоска	0	1	8	61	2	3	9	16
разочарование	0	4	41	1	0	17	26	9
отчаяние	0	20	4	0	8	20	46	4
уныние	7	7	0	0	0	20	40	27
горе	0	0	0	0	0	0	62	38

Частота указана в процентах от общего числа всех обнаруженных кодировок. Данные округлены до целого числа. В связи с ошибками округления сумма чисел по строке может быть не равна 100.

Легко заметить, что слова, традиционно считающиеся синонимами, обладают сходным распределением частот маркировок. Например, для слов *грусть, печаль, тоска* более частотна маркировка творительным падежом (*Он с грустью думал о предстоящем расставании*), для слов *меланхолия, хандра* – винительным с предлогом *в* (*Он опять впал в меланхолию*) и т. д. Таким образом, полученные числовые данные могут быть применены для прояснения спорных моментов при распределении слов по синонимическим рядам.

6. Методы анализа данных

К данным, представленным в табличной форме, могут быть применены различные виды математического анализа. В настоящей статье речь пойдет о трех разных методах анализа. Одним из них является кластерный анализ, который позволяет разбить множество слов на кластеры и построить из них дерево таким образом, что слова, более близкие по частоте маркировок, а значит, и по семантике, будут находиться в одном кластере и/или располагаться ближе на дереве. В [14] кластерный анализ применен к множеству слов {*грусть, печаль, тоска, уныние, меланхолия, хандра*}. В ходе анализа использовались данные из табл. 1. Результаты приведены на рис. 1.

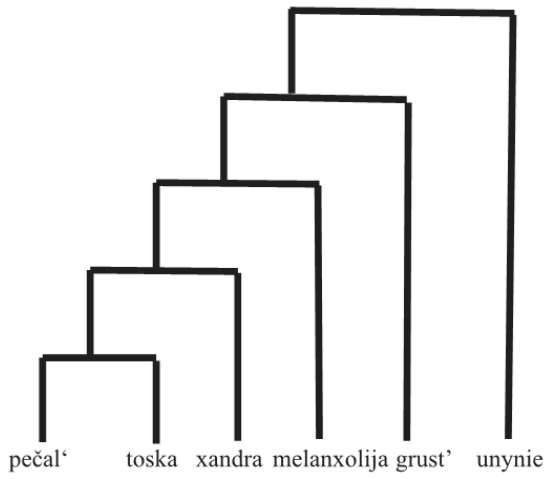

Рис. 1. Иерархическая кластерная структура (из [14])

Второй способ анализа математически близок к предыдущему, но в итоге получается не иерархическая структура, а плоская, так называемая ординация. Для построения ординации применяются специальные алгоритмы, например пакет R [15]. В результате близкие слова и на плоскости располагаются близко друг к другу. На рис. 2 ординация применена к множеству слов с общим значением *эмоции*. Полученное представление указывает на значительную сложность структуры семантического поля эмоций. В частности, здесь не выделяются явным образом центр и периферия, постулируемые когнитивной лингвистикой. Отсутствует также четкое разделение на ряды синонимов.

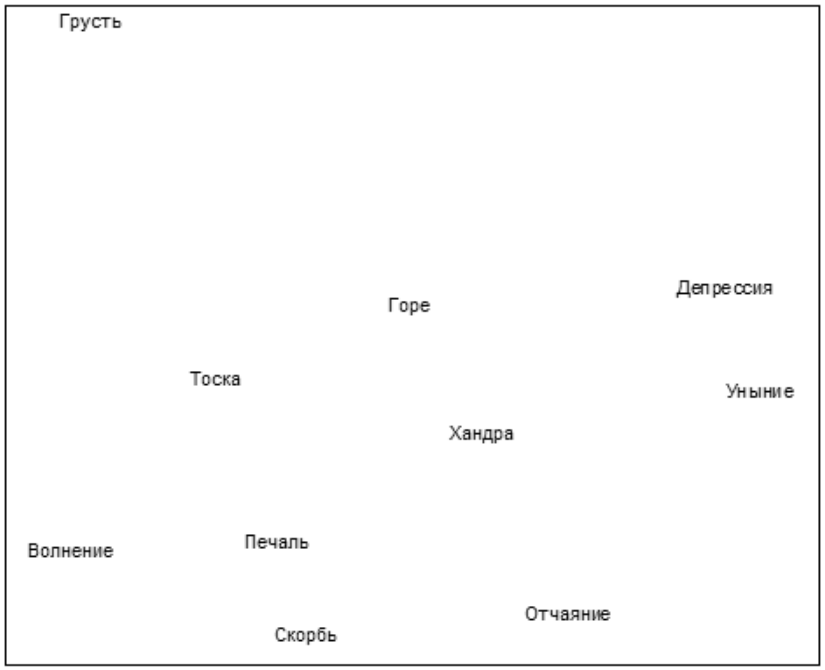

Рис. 2. Ординация множества эмоций

В наших работах [16, 17] кластерный анализ и ординация применялись к другому множеству слов – группе глаголов типа *стараться*. В качестве материала для анализа использовались как корпусные данные, так и результаты психосемантических экспериментов. Показано, что полученные результаты хорошо согласуются друг с другом.

Третий способ дает еще один вид наглядного представления данных. Впервые он был представлен на ежегодной конференции ассоциации AATSEEL славистов в Вашингтоне в 2005 г., затем обсуждался на различных международных конференциях [18, 19] и наконец был изложен в публикации [14] (на английском языке). Продемонстрируем его на примере группы слов {*ужас, отчаяние, меланхолия, депрессия*} и второго типа синтаксических отношений. Для этого построим графики частот маркировок этих слов (рис. 3). На графиках на оси абсцисс указана маркировка, на оси ординат – частота этой маркировки в процентах (из табл. 1).

Обращают на себя внимание различия в характере графиков, отражающие своеобразие указанных эмоций (точнее, психических состояний, но здесь подобное разграничение для нас не столь существенно). Можно заметить, что графики эмоций *отчаяние* и *ужас* близки. Они имеют близкие значения во всех точках. Хотя эти слова и не являются синонимами, но в их глубинной семантике есть что-то общее. С ними резко контрастирует график для слова *депрессия*, чего и следовало ожидать по причине больших различий в семантике. В отличие от эмоций *отчаяние* и *ужас*, *депрессия* имеет ярко выраженный пик в точке 2 (вин. пад. + *в*) и впадины в точках 1 (пред. пад. + *в*) и 4 (твор. пад. + *с*). График для слова меланхолия явно ближе к графику для депрессии (эти слова, видимо, почти синонимы), хотя пик в точке 2 не столь ярко выражен.

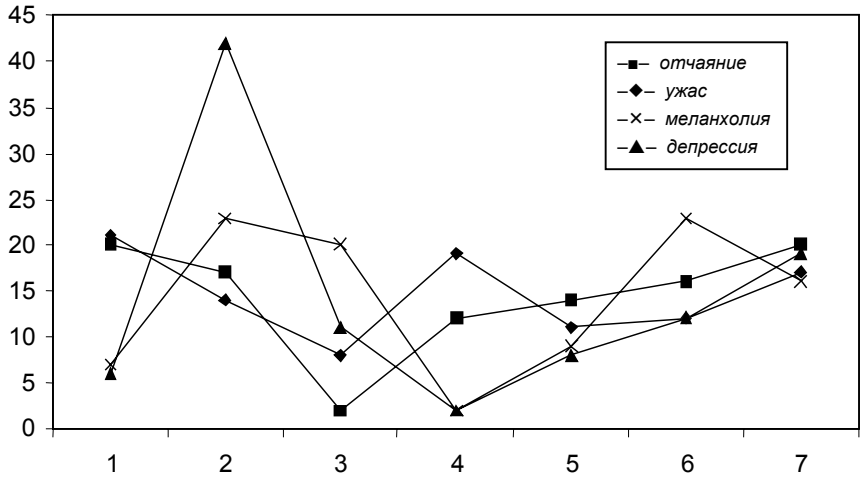

Рис. 3. Графики синонимов *ужас*, *отчаяние*, *меланхолия*, *депрессия*. Ось *X*: 1 – пред + *в*, 2 – вин + *в*, 3 – твор, 4 – твор + *с*, 5 – род + *от*, 6 – вин, 7 – остальные

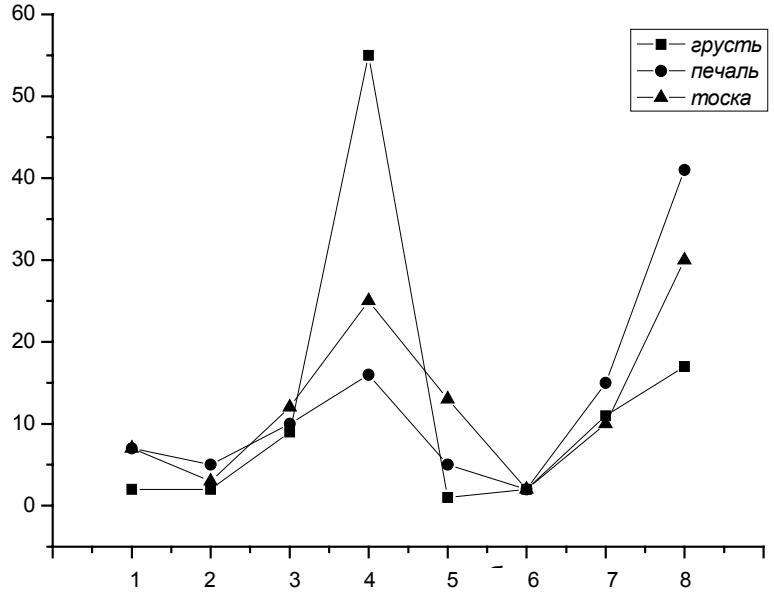

Рис. 4. Графики синонимов *грусть*, *тоска*, *печаль*. Ось *X*: 1 – вин + *в*, 2 – вин + *в*, 3 – твор, 4 – твор + *с*, 5 – род + *от*, 6 – род + *без*, вин, 7 – остальные, 8 – вин

Приведем графики слов, обычно признаваемых синонимами: *грусть*, *тоска*, *печаль* (рис. 4).

Как видим, графики на рис. 4 имеют одинаковую форму, отличаясь только величиной пиков. Таким образом, идентичная структура графиков коррелирует с близостью слов по смыслу.

При сравнении этого метода с кластерным анализом и ординацией следует отметить некоторые особенности. При применении кластерного анализа и ординации учитываются данные для всех рассматриваемых маркировок без особого выделения какой-либо из них. При сопоставлении же графиков на диаграмме

обращает на себя внимание общая форма графиков – наличие пиков и впадин и места их расположения. Точные значения частот маркировок, не относящихся к пикам, не играют существенной роли. Таким образом, фактически сравнение проводится по двум-трем выделенным параметрам, что можно охарактеризовать как качественный анализ количественных данных.

Проанализируем причины эффективности этого способа, то есть возможность выделения небольшого числа ключевых параметров. Вернемся к примеру со словом *стул.* Там тоже лишь три из большого числа возможных маркировок оказываются частотными. Пользуясь терминологией, характерной для когнитивной лингвистики и теории метафоры, можно сказать, что эти маркировки отражают отнесение стула к классу поверхностей, который рассматривается в одном ряду с контейнерами и другими онтологическими категориями. Для каждой онтологической категории есть характерные способы использования объектов этой категории в человеческой деятельности. Так, поверхности чаще всего используются следующим образом: на них что-либо кладут, это что-то там находится и его оттуда убирают. Таким образом, частотные маркировки, отражая основные способы использования объекта, относят его к одному из онтологических классов.

Данные наблюдения отличают развиваемый здесь подход от корпусного анализа семантики в [3]. В этой работе рассматривалась проблема описания синонимии в некотором семантическом поле глаголов. Было выбрано 87 параметров описания глаголов – вероятно, практически все, которые можно извлекать из корпуса более или менее формальным образом. В нашем подходе используется значительно меньшее число параметров.

Прежде всего основные онтологические свойства существительных выражаются с помощью падежно-предложных маркировок дополнений и обстоятельств, так что достаточно рассмотреть их. Очевидно, что в каждом семантическом поле (мы это демонстрируем на примере поля эмоций) из всех маркировок, как правило, используется лишь небольшая часть. То есть достаточно проанализировать частоту встречаемости менее чем 10 основных маркировок, чтобы получить основную информацию о лексеме. Более того, как показано выше, за эмоциями (и словом *стул*) стоят лишь 1–2 онтологические метафоры, определяющие и объясняющие основные употребления лексемы. Характерным является и пример с анализом слова *страх* у Апресяна, в котором в 12 из 15 рассмотренных конструкций слово *страх* имеет одну единственную маркировку род + *от.* В работе Н.Д. Арутюновой [11] метафора эмоций как жидкости обосновывается с помощью 30 примеров с падежно-предложными маркировками, из которых 26 относятся к 6 наиболее частотным, выделенным в настоящей работе.

Заключение

С появлением представительных корпусов текстов появилась возможность проведения семантических исследований, не опирающихся на интуицию отдельных исследователей. Получаемые с помощью корпусов и математической обработки данных результаты являются более объективными и приближают по строгости используемых методов исходно гуманитарную лингвистику к естественным наукам.

Наш подход основан на учете синтаксических отношений. Ввиду отсутствия хороших синтаксических анализаторов разбор предложений проводился вручную. Выбор простого множества четко маркируемых на морфологическом уровне синтаксических отношений позволил практически исключить возможность ошибок и влияния субъективного мнения исследователей.

Сокращение числа обсчитываемых параметров позволяет значительно уменьшить трудоемкость метода. Кроме того, следует обратить внимание на то, что все параметры (предложно-падежные маркировки) явно выражаются на уровне поверхностного синтаксиса (морфологии). Их извлечение не требует никакой интуиции исследователя и может быть автоматизировано, например, при наличии хорошего синтаксического анализатора.

Описанный в статье метод позволяет представить четкие и недвусмысленные результаты по проблеме синонимии. Предложено новое определение синонимии как близости синтаксических свойств слов. На конкретных примерах показано, как работает данный метод.

Предложенный метод имеет следующую когнитивную основу: каждый концепт, обозначаемый существительным, вероятно, относится к небольшому числу ключевых онтологических категорий; основные же онтологические категории проявляют себя на поверхностном уровне через определенные морфологические маркировки. Таким образом, выделение наиболее частотных морфологических маркировок, которые встречаются с исследуемым словом, позволяет определить его онтологический статус, являющийся ключевым при описании семантики.

Работа выполнена при финансовой поддержке РФФИ (проект № 07-06-00221а.)

Summary

V.D. Solovyev, V.R. Bayrasheva. Morphosyntactic Method for Semantic Information Extraction from Corpora.

The article is devoted to the description of a new method, which allows the researcher to extract new information about semantics of the words from corpus texts. The method is based on the separation of syntactic relations within the word under investigation and the text, on the definition of relations' morphological marking and receiving the frequency characteristics of marking results. The method is directed to the extraction of new information from different types of corpora, using only formal signs, which can be easily and accurately distinguished without the researcher's semantic intuition. The abilities of using this method are demonstrated on the example of determination the problem of synonymy. Results, which have been received earlier, are summarized and new data are given. This method can probably be used in the description of metaphors and semantic fields' structure.

Key words: corpora, semantics, statistic analysis, synonymy, vocabulary of *emotions* semantic field.

Литература

1. *Deignan A.* Metaphor and Corpus Linguistics. – Amsterdam: John Benjamins, 2005. – 235 p.
2. Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М.: Индрик, 2005. – 344 с.

3. *Divjak D., Gries S.* Ways of trying in Russian: clustering behavioral profiles // Corpus Linguistics and Linguistic Theory. – 2006. – V. 2, No 1. – P. 23–60.
4. *Landauer T.K., Foltz P.W., Laham D.* Introduction to Latent Semantic Analysis // Discourse Processes. – 1998. – V. 25. – P. 259–284.
5. *Апресян Ю.Д. и др.* Новый объяснительный словарь синонимов русского языка. Вып. 1. – М.: Языки рус. культуры. 1997. – 512 с.
6. *Александрова Е.* Словарь синонимов русского языка. – М., 1989. – 568 с.
7. Словарь синонимов русского языка / Ред. А.П. Евгеньева. – М.: АСТ: Астрель, 2002. – 648 с.
8. Толковый словарь русского языка / Ред. Д.Н. Ушаков. – М. АСТ: Астрель, 2000. – 1424 с.
9. *Ortony A., Clore G., Collins A.* The Cognitive Structure of Emotions. – N. Y.: Cambridge Univ. Press, 1988. – 207 p.
10. *Апресян Ю.Д.* Метафора в семантическом представлении эмоций // Вопр. языкознания. – 1993. – № 3. – С. 27–35.
11. *Арутюнова Н.Д.* Предложение и его смысл. – М., 1976. – 383 с.
12. *Апресян Ю.Д.* Экспериментальное исследование семантики русского глагола. – М.: Наука, 1967. – 251 с.
13. *Апресян Ю.Д. и др.* Лингвистическое обеспечение системы ЭТАП-2. – М.: Наука, 1989. – 294 с.
14. *Janda L., Solovyev V.* What Constructional Profiles Reveal about Synonymy // Cognitive Linguistics. – 2009. – No 2. – P. 367–393.
15. *Baayen R.H.* Analyzing Linguistic Data. – Cambridge: Cambridge Univ. Press, 2008. – 353 p.
16. *Соловьев В.Д.* Методика анализа семантических репрезентаций: на пути к естественно-научной парадигме // Проблема представления (репрезентации) в языке. Типы и форматы знаний. – М.: ИЯ РАН, 2007. – С. 80–85.
17. *Соловьев В.Д., Байрашева В.Р.* О структуре семантического поля глаголов типа «стараться» // Вопр. когнитивной лингвистики. – 2007. – № 2. – С. 87–94.
18. *Janda L., Solovyev V.* Constructional profile as a measure of Synonymy/ antonymy: ‘happiness’ and ‘sadness’ in Russian. // Proc. of the Intern. Conf. “Cognitive and Functional Perspectives on Dynamic Tendencies in Languages”. – Tartu: University of Tartu, 2008. – P. 225–226.
19. *Janda L., Solovyev V.* What constructional profile reveal about synonymy and metaphor: a case study of Russian words for ‘sadness’ // Труды 3-й Междунар. конф. по когнитивной науке. – М.: Худож.-изд. Центр. 2008. – С. 74–75.
20. *Janda L., Solovyev V.* Are Emotions Metaphorical Objects? Constructional Profiles and Russian words for ‘sadness’ // Proc. of the CSDL. – Ohio: Ohio University, 2008. – P. 75.

Поступила в редакцию
18.01.10

Соловьев Валерий Дмитриевич – доктор физико-математических наук, профессор кафедры теоретической кибернетики Казанского государственного университета.

Байрашева Венера Рустамовна – кандидат физико-математических наук, доцент кафедры теоретической кибернетики Казанского государственного университета
E-mail: *maki.solovyev@mail.ru*