# О возможностях автоматизации выявления связей между терминами предметной области (на примере катализа)[1]

# Possibilities of automation of relationship identification between subject-domain terms (on the material of catalysis)

**Саломатина Н. В.** (nataly@math.nsc.ru),
**Гусев В. Д.** (gusev@math.nsc.ru)
Институт математики СО РАН, Новосибирск

**Ильина Л. Ю.** (ilud@catalysis.ru), **Кузьмин А. О.** (kuzmin@catalysis.ru),
**Пармон В. Н.**
Институт катализа СО РАН, Новосибирск

В рамках проблемы автоматизации построения тезаурусов предметных областей на базе текстовых подборок рассматриваются три подхода к выявлению связей между терминами: 1) построение профиля кластеризуемости наиболее значимых элементов текста, 2) формирование специфических шаблонов (образцов с переменными), 3) использование индикаторов связи.

## Введение

Нижние уровни *онтологий* различных предметных областей (ПО) обычно представлены *тезаурусами*, содержащими информацию об основных *понятиях и терминах* ПО, а также *связях* между ними. Автоматизация построения онтологий *на основе текстов* ПО является актуальной задачей компьютерной лингвистики. Для формирования терминологических словарей разработано довольно много компьютерных подходов, минимизирующих затраты ручного труда [1÷4]. Гораздо меньше работ посвящено *выявлению связей* между терминами ПО [5, 6]. *Целью* данной работы является расширение спектра возможных подходов к *автоматизации* (хотя бы частичной) этого процесса.

*Отличительной особенностью* предлагаемых подходов является использование техники $L$-граммного анализа в сочетании с позиционным как на этапе формирования словаря по текстам ПО, так и на этапе выявления связей между его элементами. Термин $L$-грамма, по-видимому, был впервые введен Шенноном в [7] применительно к цепочке из $L$ подряд следующих букв текста, а затем был перенесен (не совсем корректно) и на цепочки из $L$ подряд следующих слов ($L$ = 1, 2, …). Именно в последнем смысле он используется в данной работе. Техника $L$-граммного представления разработана нами как применительно к одному, так и к группе текстов [8]. В первом случае с ее помощью выявляются всевозможные внутритекстовые повторы произвольной длины, во втором случае — межтекстовые повторы, что удобно для целей классификации. Наряду с информацией о частоте встречаемости каждой $L$-граммы в тексте, фиксируются места ее вхождения в текст (позиционная информация).

Привлекательными особенностями $L$-граммного подхода к формированию тезаурусов ПО являются: применимость к разноязычным текстам, ориентация на извлечение терминов произвольной длины, оценка их информативности путем привлечения позиционной информации, возможность формирования шаблонов для описания групп близких $L$-грамм и установления связей между ними. Важно отметить, что $L$-граммные спектры содержат не только терминологические цепочки, но и индикаторные, несущие информацию о связях между терминами.

[1] Работа выполнена при финансовой поддержке Интеграционного проекта СО РАН № 111.

## 1. Исходные данные и предобработка

Предметная область, используемая нами для иллюстрации предлагаемых подходов, связана с разделом химии, изучающим возможности ускорения или замедления химических реакций (катализ). Исходная подборка была представлена пятью текстами: z1: О. В. Крылов «Гетерогенный катализ» (учебник); z2: В. Б. Фенелонов «Введение в основы адсорбции и текстурологии» (учебник); z3: И. П. Мухленов «Технология катализаторов»; z4: «Лекции по катализу» (1÷15); z5: «Химическая энциклопедия» (фрагменты). Суммарный объем подборки — свыше 403 тыс. словоупотреблений. Характерные особенности подборки: наличие значительного числа синонимов, связанных с дублированием названий веществ их химическими формулами; большое число аббревиатур и сокращений, в том числе общеупотребительных слов (см. z5); вариативность в числе слов, используемых для обозначения одного и того же понятия (метанол ($L = 1$) ≡ метиловый спирт ($L = 2$)); наличие специфических терминов в каждом из источников (z1÷z5) (и только в нем), лежащих на периферии основной проблематики. Все эти факторы в значительной степени влияют на результаты и должны учитываться при автоматической обработке.

*Предобработка* исходных материалов состояла из следующих этапов:

- Нормализация текстовой подборки;
- Получение $L$-граммных характеристик [8] всей подборки для значений $L = 1, 2, \ldots, L_{max}$, где $L_{max}$ — длина (число слов) максимальной повторяющейся цепочки в нормализованной подборке. В характеристике $L$-го порядка представлен полный спектр $L$-грамм, присутствующих в подборке, с указанием их частот встречаемости и распределения по отдельным источникам z1÷z5.
- Упорядочение $L$-граммных спектров при каждом значении $L$: а) по убыванию частоты встречаемости; б) лексикографически; в) по убыванию показателя неравномерности позиционного распределения (более детально об использовании этого показателя см. в [15]).

Указанные этапы являются общими как при формировании словаря, так и при выявлении связей между его элементами. Последующие этапы могут различаться в зависимости от преследуемой цели, но все они носят характер процедур *фильтрации* для отсеивания малоинформативной (в интересующем нас плане) части $L$-граммного спектра. Примерами процедур фильтрации являются: — отбор $L$-грамм ($L \geq 2$), удовлетворяющих *критерию устойчивости* [2]. Это основа для выделения многословных терминов и индикаторов связи;[2] — *учет частотной и позиционной информации* (отсеиваются низкочастотные $L$-граммы и $L$-граммы с равномерным позиционным распределением); — проверка наличия синтаксической связности слов в цепочке (желательна для элементов терминологического словаря, но необязательна для индикаторов связи); — учет частеречных значений (в терминологических сочетаниях преобладают существительные и прилагательные, среди индикаторов связи встречаются и глаголы) и др.

Заметим, что упомянутые выше процедуры упорядочения тоже, в некотором смысле, можно трактовать как процедуры фильтрации, позволяющие провести отсечение «избыточного» материала в нужном месте. Так, упорядочение в) призвано сдвинуть вниз $L$-граммы общеупотребительного толка, распределенные, как правило, равномерно по тексту. Это способствует концентрации терминоподобных $L$-грамм в начальной части списка. Наибольший интерес представляют $L$-граммы, поднявшиеся вверх в упорядочении в) по сравнению с упорядочением а). Например, в упорядочении а) слова *катализатор*, *поверхность*, *реакция*, *адсорбция* занимали соответственно, 6-е, 10-е, 15-е и 22-е место. В упорядочении в) они поднялись, соответственно, на 1-е, 5-е, 4-е и 2-е место. Таким образом, работая с упорядочением в) эксперт может существенно уменьшить объем просматриваемого материала.

## 2. Возможные подходы к выявлению связей между терминами

Наибольшую трудность представляет автоматизация процедуры выявления связей между элементами терминологического словаря и уточнение номенклатуры этих связей. На данный момент мы рассматриваем три возможности продвижения в этом направлении, связанные с использованием: а) профилей кластеризуемости; б) терминологических шаблонов; в) индикаторов связи.

### 2.1. Профили кластеризуемости наиболее значимых элементов текста

Этот аппарат ориентирован в первую очередь на выявление *ассоциативных связей* между элемен-

[2] Термином *устойчивая цепочка* мы характеризуем $L$-граммы ($L \geq 2$), встречающиеся в большом числе разнообразных контекстов. И, наоборот, неустойчивой считается цепочка, которая лишь единственным образом продолжается вправо или влево при всех своих вхождениях в текст. Это означает, что она не имеет самостоятельного значения и функционирует лишь в составе одной и той же (более длинной) цепочки. Детали формализации понятия *устойчивости* описаны в [8].

тами словаря. В отличие от совместной встречаемости, предполагающей позиционную близость двух (или большего количества) слов в рамках какого-либо устойчивого словосочетания, ассоциативно связанные слова (или словосочетания) могут быть разнесены друг от друга в общем случае на произвольное расстояние. Предполагается, что начальная версия словаря уже получена и представлена (в большинстве своем) $L$-граммами ($L \geq 1$), демонстрирующими неравномерное распределение в тексте. Наиболее характерное проявление неравномерности связано с кластеризацией вхождений $L$-граммы в отдельных участках текста. Статистически значимые кластеры могут быть выделены с помощью сканирующих статистик [9]. Кластеры, образуемые разными $L$-граммами, могут быть позиционно разнесены друг от друга, пересекаться друг с другом или вкладываться один в другой. Понятие *профиля кластеризуемости* было введено нами в [10], чтобы аккумулировать на одном графике информацию обо всех участках кластеризации разных $L$-грамм.

Формально, профиль кластеризуемости — это ступенчатая функция, аргументом которой является порядковый номер предложения в тексте, а значение равно числу различных кластеров, включающих в себя данное предложение. При этом в рассматриваемом предложении вовсе не обязаны присутствовать одновременно все $L$-граммы, кластеризующиеся в данном участке текста. Пики профиля кластеризуемости обычно соответствуют отдельным микротемам текста, а провалы между ними — переходу от одной микротемы к другой. Использование профилей кластеризуемости для выявления ассоциативных связей основано на предположении о том, что такого рода связи как раз и имеют место между элементами словаря, кластеризующимися в одном и том же участке текста.

Ниже на *Схеме 1* приведен фрагмент профиля кластеризуемости текста z1, охватывающий предложения с номерами от 3973 до 5287. Здесь ось абсцисс с номерами предложений направлена вниз, а ось ординат (число кластеров) — по горизонтали:

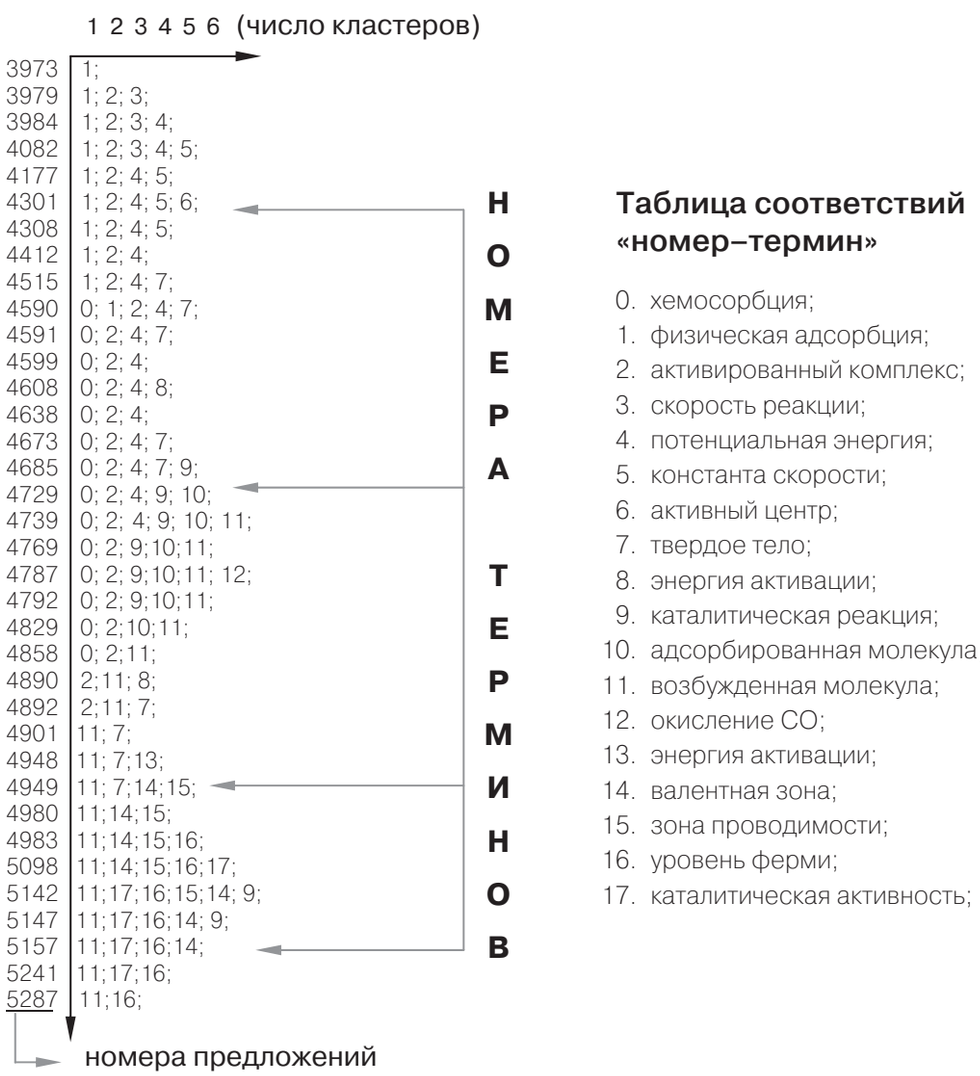


**Схема 1.** Профиль кластеризуемости фрагмента текста z1

слева направо. Для экономии места ось ординат представлена в нелинейном масштабе: указаны номера лишь тех предложений, на которых происходит изменение значений профиля, т. е. добавляются новые кластеры или исчезают старые. Относительно редкие случаи сохранения высоты соседних столбиков связаны с одновременным добавлением и устранением кластера (см., например, №№ 4515 и 4591).

Для наглядности профиль в каждой точке представлен набором чисел, отделенных друг от друга точкой с запятой. Количество чисел в наборе (значение профиля) соответствует числу кластеров, включающих в себя данное предложение. Сами же числа — это номера элементов словаря, по которым строился профиль. Таблица соответствий «номер–термин» представлена рядом с графиком. Опять же для упрощения картины профиль строился лишь по 80 биграммным комбинациям (упорядочение в)). Из них на рассматриваемом участке проявили себя в виде позиционных кластеров лишь 18[3].

В принципе, каждый набор чисел, соответствующий конкретной позиции текста, можно трактовать как набор *ассоциативно связанных терминов*. Нетрудно видеть, что соседние наборы сильно коррелируют друг с другом. Для позиционно разнесенных наборов корреляция падает. Например 6-ти элементные наборы, соответствующие предложениям с номерами 4787 и 5142 имеют лишь два общих термина: *каталитическая реакция* и *возбужденная молекула* (№№ 9 и 11 в таблице соответствий).

Наивысшее значение профиля на тексте z1 (длина 17859 предложений) равно 8 и зафиксировано в предложениях с номерами 2782, 3004 и 13202. Приведем для иллюстрации список терминов, кластеризующихся в районе позиции 13202: *каталитическая реакция*, *каталитическая активность*, *активный центр*, *атом металла*, *число атомов*, *частица металла*, *размер частицы*, *адсорбция* $H_2$. Нетрудно видеть, что ассоциативные связи между отдельными парами терминов проявляют себя в том числе и на уровне общих словоформ.

Сила ассоциативной связи между любой парой терминов, по-видимому, может характеризоваться числом наборов, в которых они совместно встречаются, позиционной привязкой этих наборов (соседние или разнесенные) и, возможно, другими факторами. Вопрос требует специального изучения с привлечением экспертов ПО. Заметим также, что использование ассоциативных связей для повышения эффективности информационного поиска не всегда приводит к успеху. Тем не менее в существующих стандартах на построение тезаурусов различных предметных областей этот тип связей фигурирует.

[3] Термин «хемосорбция» формально является одногаммой, но фактически это биграмма (химическая адсорбция).

## 2.2. Терминологические шаблоны (образцы с переменными)

Понятие шаблона используется в различных языковых системах и подразумевает группу символьных объектов, объединенных в один класс по какому-то общему структурному признаку. Шаблоны многолики. Например, для описания регуляторных фрагментов в генетических текстах используют частично-специфицированные строки символов или строки с элементами типа “don’t care”. В качестве образца может выступать регулярное выражение. Поиск по регулярному выражению реализован, например, в системе ALEX [11]. Лексико-синтаксические шаблоны, предназначенные для распознавания специфических языковых конструкций (например, согласованных именных словосочетаний), обсуждаются в [12].

Применительно к используемой нами *L*-граммной системе представления текстов нас будут интересовать шаблоны, объединяющие в один класс сходные (отличающиеся не более, чем по 1–2 позициям) цепочки слов. Формально, такого типа шаблоны можно рассматривать как частный случай образцов с переменными (см. [13]), где переменные указывают на позиции, допускающие варьирование. Например, шаблон из 3-х слов с одной переменной *x*, имеющий форму *p = производство x кислоты*, допускает вместо *x* следующие подстановки: *серной* (встречается в исходной подборке 11 раз), *азотной* (3 раза), *пропионовой* (1), *акриловой* (1). Эти термины связаны друг с другом отношением принадлежности к одному таксону «типы кислот». Образец с двумя переменными (*x* и *y*), имеющий форму *p = окисление x в y*, допускает следующие пары согласованных подстановок: *x = этилена, y = этиленоксид* (эта пара встретилась 11 раз); *x = пропилена, y = акролеин* (16 раз); *x = метана, y = метанол* (1 раз) и др. Пары *x, y* связаны участием в одном процессе (окисление).

Связи типа «общее–частное» часто выявляются путем установления факта вложения одной *L*-граммы в другую (более длинную)[4]. Например, если в словаре имеется термин $p_1$ = *окисление этилена*, то поиск по образцу $p_2$ = *окисление этилена в x* выявит термины, являющиеся более узкими по отношению к $p_1$ (например, *окисление этилена в этиленоксид* или *окисление этилена в ацетальдегид*). Аналогично, сужением термина *кислотный центр* в соответствии с образцом *p = кислотный центр x* будут термины *кислотный цент Льюиса* и кислотный *центр Бренстеда*.

Формирование образцов с одной переменной осуществляется очень просто. Пусть, для примера,

[4] Все приводимые в данном абзаце примеры вложений касаются пар цепочек, относимых экспертами к терминам предметной области.

$L = 3$. Вычисляем по исходным текстам $L$-граммную характеристику 3-го порядка, содержащую полный спектр $L$-грамм, представленных в тексте, с указанием их частот. Заменяем словоформы, стоящие в первой позиции каждой 3-граммы, элементом $x$. При этом «склеиваются» (становятся неразличимыми) все 3-граммы, отличавшиеся только по первой позиции, и возникает множество образцов вида $p = x\ a\ в$, где $a$ и $в$ — фиксированные канонические формы, например, $p = x$ *активированный комлекс*, где x ∈ {образование, *конфигурация*, *модель*, *теория*,…}. Аналогично, заменяем элементом $x$ словоформы, стоящие во 2-й позиции каждой 3-граммы. При этом склеиваются все триграммы, отличающиеся только по этой позиции, и возникает множество образцов вида $p = a\ x\ в$, где $a$ и $в$ — фиксированные канонические формы, например, $p =$ *образование* $x$ *комплекс*, где $x \in$ {*активированный*, *поверхностный*, *мультиплетный*, *сверхкислотный*, *сульфитный*, *низкоспиновый*, *высокоспиновый*,…}. Здесь список допустимых значений содержит перечисление типов комплексов и указывает на наличие антонимических связей между ними (*низкоспиновый* → *высокоспиновый*). Наконец, осуществляя подстановку $x$ по третьей позиции, получаем множество образцов вида $p = a\ в\ x$. Две переменные имеет смысл вводить лишь для длинных $L$-грамм ($L \geq 4$).

Как показывает приведенный выше пример, анализ допустимых подстановок в образцах дает важную информацию о предполагаемых связях между объектами. Возникающие при этом трудности проиллюстрируем на примере образца $p = x$ *реактор*, допускающего более 100 вариантов различных подстановок в качестве значения переменной $x$ (на исходном материале). Среди этих подстановок можно выделить группу, характеризующую типы реакторов: *каталитический* (встретился в текстах 21 раз), *проточный* (13), *трубчатый* (9), *адиабатический* (9), *изотермический* (5) и др. Она представляет основной интерес. Другая группа подстановок характеризует конструктивные особенности реактора: *конструкция* (3), *корпус* (4), *центр* (3), *освинцованный* (1), *железный* (1). Третья группа носит характер «шума»: *промышленный* (9), *пустой* (3), *распространенный* (2), *третий* (1), *изнутри* (1), *рассчитывать* (1). Нетрудно видеть, что для выделения интересующей нас группы, важной является информация о частеречных значениях и, в меньшей степени, о частоте встречаемости. Возможность фильтрации по этим параметрам предусмотрена в программе. Но даже если устранены числительные, глаголы и наречия, а также однократно встречающиеся объекты, эксперту придется разбираться со случаями типа *промышленный* (9), *пустой* (3) и т. п.

Завершая этот раздел, заметим, что формирование образцов и их последующий анализ напоминает изучение конкордансов, собранных вместе «на все случаи жизни». Некоторая избыточность такого подхода, тем не менее, оправдана, поскольку образцы несут в себе элемент обобщения. Задавая, например, поисковый запрос в виде *окисление x в y*, мы получим не только пары *x* и *y*, фигурировавшие в исходной подборке но и многие другие в ней отсутствовавшие.

### 2.3. Индикаторы связи

Понятие индикатора того или иного аспекта содержания текста известно давно (см. обзор [14]). Индикаторы могут использоваться и для обнаружения в тексте упоминаний о каких-либо объектах, событиях и т. п. Применительно к интересующей нас задаче (формирование тезауруса ПО) индикаторы могут быть использованы для выявления связей между понятиями ПО. К достоинствам индикаторного подхода следует отнести простоту реализации, интерпретируемость результатов, к ограничениям — необходимость формирования индикаторных словарей в каждом отдельном случае (как правило, вручную) и отсутствие гарантий обязательного присутствия индикатора.

Иллюстрирующим примером наличия индикаторов связи в тексте может служить фраза из раздела z5 подборки, принадлежащая О.В. Крылову: «Он (Ипатьев)… создал ряд важнейших каталитических процессов нефтепереработки, *таких как* алкилирование, гидрокрекинг, изомеризация». Здесь индикатором связи «общее–частное» выступает биграмма *таких как*, роль «общего» играет выделенный левый контекст этой биграммы, а «частного» — правый.

Индикаторы связи не являются элементами терминологического словаря, но отбираются параллельно с его формированием путем просмотра экспертом устойчивых цепочек ($L$-грамм), упорядоченных по убыванию показателя неравномерности позиционного распределения. В отличие от терминов ПО индикатор может даже не удовлетворять требованию синтаксической связности. Таковым, например, является индикатор причинно-следственной связи *приводить к* (частота встречаемости в подборке $F = 278$).

Кроме уже упомянутых индикаторов *такой как*, *приводить к* было выделено еще несколько десятков индикаторов связи разного типа. Для их отбора эксперту пришлось просмотреть около 2000 устойчивых двухсловных сочетаний с частотой встречаемости $F \geq 10$. Напомним, что для выделения их непосредственно из текстовой подборки нужно было бы прочитать порядка 400 тыс. слов. Укажем некоторые из отобранных индикаторов: — *и др.* ($F = 284$, типы связи — «общее–частное», принадлежность к одному таксономическому классу). Текстовый пример: «В гомогенном кислотном катализе *в качестве* катализаторов используют протонные кислоты ($H_2SO_4$, $HC_1$, $H_3PO_4$ *и др.*)». Здесь индикатор *и др.* связывает

каждую из конкретных кислот с обобщающим термином *протонные кислоты*. Кроме этого имеется еще один индикатор *в качестве*, который связывает термины *катализатор* и *протонные кислоты*. Заметим, что индикатор *и др.* срабатывает практически без ошибок; — *один из* ($F = 180$, типы связи: «общее–частное», ассоциативная). Текстовый пример 1: «В нефтепереработке алкилирование используется как *один из* методов повышения октанового числа бензина». Здесь *алкилирование* ассоциативно связывается с *октановым числом*. Текстовый пример 2: «Если *один из* реагентов связывается сильно, а другой распределен равномерно между обеими фазами, то …». Этот пример иллюстрирует ситуацию, когда индикатор не срабатывает.

Кроме рассмотренных можно упомянуть такие индикаторы как *в том числе* ($F = 16$), *использовать в качестве* ($F = 16$, условная синонимия), *представлять собой* ($F = 82$, часто используется как элемент определения), *состоящий из* ($F = 72$, часть–целое), *связанный с* ($F = 136$) и др. Как уже было показано на текстовом примере 2, индикаторы не всегда фиксируют связь, но их значительное разнообразие и высокая частота встречаемости в тексте позволяют выявить множество связей в формируемом тезаурусе. Заметим также, что многие из перечисленных индикаторов можно трактовать и как маркеры фактов (например, *приводит к*) или элементы определений (*так называемый*, *представлять собой* и др.), что расширяет сферу применимости индикаторного подхода.

## Заключение

Предложены три возможных подхода к выявлению связей между элементами тезауруса предметной области (катализ), формируемого на основе анализа достаточно представительной текстовой подборки. Рассмотрены возможности автоматизации (частичной) этого процесса в рамках используемой авторами *L*-граммной системы представления текстов. Они ориентированы на минимизацию неизбежного (на данный момент) ручного труда эксперта на заключительном этапе.

## Литература


1. *Dobrov B., Loukachevitch N., Nevzorova O.* An approach to new ontologies development: main ideas and simulation results // Int. J. Information Theories & Applications. — Vol. 10, N 1, 2003. — P. 98–105.
2. *Гусев В. Д., Саломатина Н. В.* Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог–2004), М.: Наука, 2004. — С. 530–535.
3. *Гельбух А. Ф., Сидоров Г. О., Эрнандес-Рубио Э., Чубукова М. В.* Словари сочетаемости слов: какой метод составления лучше? // Там же. — С. 133–138.
4. *Браславский П. И., Соколов Е. А.* Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2006), М.: Изд. РГГУ, 2006. — С. 88–94.
5. *Кузнецов П. И.* Лингвистические и алгоритмические аспекты выделения объектов и связей из предметно-ориентированных текстов // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2007), М.: Изд. РГГУ, 2007. — С. 333–342.
6. *Шабанов В. И., Власова А. Е.* Алгоритм формирования ассоциативных связей и его применение в поисковых системах // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2003), М.: Наука, 2003. — С. 603–608.
7. *Шеннон К.* Предсказание и энтропия печатного английского текста // Работы по теории информации и кибернетике. — М.: Изд. ИЛ, 1963. — С. 669–686.
8. *Гусев В. Д., Саломатина Н. В.* L-граммное представление текстов на естественном языке и его возможности // Материалы Всерос. научн. конф. «Квантитативная лингвистика: исследования и модели» (КЛИМ–2005). — Новосибирск: Изд. НГПУ, 2005. — С. 256-270.
9. *Гусев В. Д., Немытикова Л. А., Саломатина Н. В.* Выявление аномалий в распределении слов или связных цепочек символов по длине текста // Вычислительные системы, вып. 171. — Новосибирск, ИМ СО РАН, 2002. — С.51–74.
10. *Гусев В.Д, Мирошниченко Л. А., Саломатина Н. В.* Профиль кластеризуемости текстов и возможности его использования // MegaLing2006. Горизонти прикладної лінгвістіки та лінгвістичних технологій. Доповіді міжнародної конференції. — Сімферополь: Вид-во «ДиАйПи», 2006. — С. 203–204.
11. *Жигалов В. А., Жигалов Д. В., Жуков А. А., Кононенко И. С., Соколова Е. Г., Толдова С. Ю.* Система ALEX как средство многоцелевой автоматизированной обработки текстов // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог–2002), М.: Наука, 2002. — С. 192–208.
12. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2007), М.: Изд. РГГУ, 2007. — С. 70–75.
13. *Handbook* of Formal Languages // G. Rosenberg, A. Salomaa (Eds), Vol.1, 1996. — Ch.4.
14. *Пащенко Н. А., Кнорина Л. В., Молчанова Т. В.* и др. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика, т. 7. — 1983 г. — С. 7–164.
15. *Остапенко В. А.* Выделение и классификация терминов с помощью элементарных квантитативных моделей // НТИ, сер. 2, № 11, 1989. — С. 24–28.