РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 10 Москва 2014

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 510.6 : 004.89

М.И. Забежайло

Приближенный ДСМ-метод на примерах

Обсуждаются возможности оптимизации перебора при интеллектуальном анализе данных средствами ДСМ-метода автоматического формирования гипотез. Понятие приближенного ДСМ-метода, его наиболее существенные алгоритмические характеристики и особенности разбираются на примерах.

Демонстрируются некоторые дополнительные возможности эффективной обработки больших коллекций эмпирических данных за счет комбинирования ДСМ-анализа со статистическим методами восстановления зависимостей из данных.

Ключевые слова: *ДСМ-метод автоматического порождения гипотез, вычислительная сложность и оптимизация перебора, приближенный ДСМ-метод*

ДСМ-метод (см., например, [1, 2] и др.) в его текущем состоянии развития – достаточно интересная платформа для ведения интеллектуального анализа данных (ИАД). Однако в ряде практических приложений его полноформатное применение в общем случае оказывается ограниченным известными (экспоненциальными) оценками вычислительной сложности переборных задач, характерных для этой технологии ИАД. Проблему объема вычислений при использовании ДСМ-ИАД в приложениях можно трактовать как своего рода «проклятие размерности», ведь классическая версия ДСМ-метода реализует *исчерпывающий* комбинаторный перебор сходств, порождаемых из заданного для обучения набора эмпирических данных (множества исходных примеров и контрпримеров).

Известно несколько подходов, используемых на практике для борьбы с названным выше «проклятием размерности» – экспоненциальной вычислительной сложностью алгоритмов, реализующих определенные компоненты ДСМ-метода. Так, например, в работах [3-5] демонстрируется, что в ряде случаев практически полезным оказывается использование соответствующих структурных фильтров (проблемно-ориентированных ограничений на структуру порождаемых ДСМ-зависимостей), а в работах [6, 7] –

высокая эффективность применения в ДСМ-расчетах весьма изощренной техники генетических алгоритмов.

В работах [8, 9] представлена техника оптимизации перебора в рамках так называемого *приближенного* ДСМ-метода, дающего возможность управлять объемами вычислений, которые необходимы для реализации ДСМ-рассуждений. В основе такой оптимизации лежит специальным образом организованная технология гибкой настройки «навигации» в множестве соответствующих ДСМ-сходств (целенаправленное управление перебором элементов множества потенциально порождаемых в каждом конкретном случае ДСМ-гипотез). Таким образом сформирована технология, позволяющая вести средствами ДСМ-метода интеллектуальный анализ данных, в том числе и на больших массивах исходных данных.

Говоря о сфере приложений представленной технологии, обратимся в первую очередь к тем областям, где от применения ДСМ-ИАД можно было бы ожидать интересных и важных прикладных результатов. Практически значимые примеры подобной проблематики дают задачи выявления, анализа и устранения сбоев в сложных программно-технических комплексах (так называемые Hi-Techdiagnostics&troubleshooting), в том числе – это задачи:

- управления ИТ-ресурсами и сервисами (ITRM\ITSM) в условиях использования технологий виртуализации вычислительных ресурсов, ресурсов хранения данных, а также – сетевых ресурсов;
- мониторинга и балансировки (оптимизации) нагрузки в крупных центрах обработки данных (ЦОДах);
- отладки (обеспечения корректного функционирования) крупных компьютерных сетей (в том числе – контроль достижимости из *заданных* входных точек пакетами *заданного* вида *заданных* выходных точек, выявление и удаление петель, контроль разделения слайсов и т.п.);
- обеспечения информационной безопасности (мониторинг и защита от несанкционированных вторжений, идентификация и пресечение атак и др.),

а также ряд других задач (в том числе – проактивной) диагностики и обеспечения бесперебойной работы высокотехнологичного компьютерного оборудования (сложных программно-технических комплексов).

Выделим отличительные особенности прикладных задач подобного типа:

- необходимость оперировать в режиме реального времени (очень жесткие ограничения по времени анализа данных, принятия и исполнения соответствующих решений);

- жесткие требования по линии SLA[1] с бизнес-пользователями предоставляемых ИТ-сервисов;

- огромные объемы технологических данных, накапливаемых в результате мониторинга операционной деятельности соответствующих программно-технических комплексов (log'и, track'и, eventbases, …)[2].

Типовая технологическая схема работы с такими задачами выглядит примерно следующим образом:

event fixing =>
=>***incident identification*** (trouble checking) =>
=>***incident*** (problem) ***causal analysis***=>
=>***trouble shooting*** (debugging) .

Одной из критически важных проблем здесь оказывается необходимость вести *каузальный анализ сбоев*, возникающих в работе объекта мониторинга. Опыт показывает, что традиционные математические техники интеллектуального анализа данных в задачах подобного типа в подавляющем большинстве практически значимых случаев оказываются неэффективными. С одной стороны, рассуждения «в среднем» – например, результаты статистического анализа и выделения существенных корреляций между теми или иными «факторами влияния» и конечным результатом – сбоем в работе анализируемой программно-технической системы – оказываются либо слишком общими для формирования значимых заключений о поведении конкретного программно-технического комплекса в текущий момент и в текущем (фиксируемом) эксплуатационно-технологическом контексте, либо статистически малозначимыми ввиду «малого» размера выборок инцидентов (тех или иных сбоев в работе анализируемого оборудования и ПО) по сравнению с размерами фиксируемых массивов данных о поведении объекта управления. С другой стороны, дискретные методы каузального анализа, предполагающие исчерпывающий перебор вариантов в процессе выявления причинных зависимостей, объясняющих фиксируемые сбои в работе оборудования и ПО, как правило, основаны на использовании комбинаторных алгоритмов экспоненциальной вычислительной сложности. В свою очередь, именно это обстоятельство делает невозможным на практике их прямое применение на больших массивах обрабатываемых данных, характерных для существенной части реальных приложений.

Предлагаемая в настоящей работе техника интеллектуального анализа данных позволяет обойти обе эти проблемы. Во-первых, средствами ДСМ-метода реализуется исчерпывающий каузальный анализ зависимостей, характеризующих причины возникновения фиксируемых сбоев. Во-вторых, эта техника позволяет обойти ограничения, связанные с экспоненциально быстро растущим объемами перебора при поиске изучаемых (восстанавливаемых по накапливаемым эмпирическим данным о сбоях) причинных зависимостей, Действительно, за счет использования уже представленной технологии приближенных вычислений (т.е. – целенаправленно управляемого перебора вариантов) здесь можно достаточно быстро породить некоторые «полезные» (для диагностики текущей анализируемой ситуации) зависимости, а затем, если это потребуется, последовательно приближаться к ситуации, в которой оказываются исчерпывающим образом проанализированными все возможные причинные зависимости, которые можно восстановить средствами ДСМ-ИАД из эмпирических данных, накапливаемых в процессе мониторинга функционирования соответствующей программно-технической системы.

[1] Service Level Agreement – соглашение об уровне сервисов.
[2] "Это – одна из наиболее активно изучаемых областей работы с так называемыми Big Data.

* * *

Для последующих ДСМ-действий с исходно заданными прецедентами – примерами и контрпримерами – условимся, что в нашем распоряжении имеются:

- исходный алфавит $\boldsymbol{U}$ (множество образующих для анализируемых примеров)

$$\boldsymbol{U} = \{a_1, a_2, \dots, a_n\},$$

- множество (примеров) $\boldsymbol{\Omega}$-объектов (т.е. непустых множеств образующих), построенных над универсумом $\mathbf{U}$[3]:

$$\boldsymbol{\Omega} = \{A_1, A_2, \dots, A_m\} \subseteq 2^{\mathbf{U}} \setminus \varnothing .$$

Располагая множествами $\boldsymbol{U}$ и $\boldsymbol{\Omega}$, определим два отображения f и φ:

$\forall \xi \in 2^{\mathbf{U}}$ (т.е. для каждого ξ-подмножества образующих из множества $\boldsymbol{U}$)

f (ξ) = {множество всех таких $\mathbf{Y}$ из $\boldsymbol{\Omega}$, что для $\forall \mathbf{x} \in \xi$ имеет место $\mathbf{x} \in \mathbf{Y}$ для каждого из этих выбранных $\mathbf{Y}$ (т.е. это множество всех примеров из $\boldsymbol{\Omega}$, в которые все образующие $\mathbf{x}$ из заданного ξ входят одновременно)}

$\forall \zeta \in 2\Omega$ (т.е. для каждого ζ-подмножества примеров из множества $\boldsymbol{\Omega}$)

φ(ζ) = {множество всех таких $\mathbf{x}$ из $\boldsymbol{U}$, что для $\forall \mathbf{Y} \in \zeta$ имеет место $\mathbf{x} \in \mathbf{Y}$ для каждого из этих выбранных $\mathbf{x}$ (т.е. это множество всех таких образующих $\mathbf{x}$ из $\boldsymbol{U}$, которые во все примеры из заданного ζ входят одновременно)}

(при этом $2^{\mathbf{Z}}$-множество всех подмножеств множества $\mathbf{Z}$ (полагая здесь $\mathbf{Z} \in \{\boldsymbol{U}, \boldsymbol{\Omega}\}$)).

Пара отображений <f, φ> представляет собою соответствие Галуа[4], а их произведения f(φ(ζ)) и φ(f(ξ)) – соответствующие замыкания Галуа[5]: $[\_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ и $[\_]_{\boldsymbol{\Omega},\boldsymbol{U}}$. Посредством $\mathbf{GC}_{f,\varphi}(\boldsymbol{\Omega})$ и $\mathbf{GC}_{\varphi,f}(\boldsymbol{U})$ будем обозначать множества неподвижных точек соответствующих замыканий Галуа:

$$\mathbf{GC}_{f,\varphi}(\boldsymbol{\Omega}) = \{\zeta \in 2^{\boldsymbol{\Omega}} \text{ таких, что } f(\varphi(\zeta)) = \zeta\}$$

$$\mathbf{GC}_{\varphi,f}(\boldsymbol{U}) = \{\xi \in 2^{\boldsymbol{U}} \text{ таких, что } \varphi(f(\xi)) = \xi\}.$$

Каждое из этих множеств можно рассматривать как частично упорядоченное в соответствии как со взаимным вложением рассматриваемых множеств объектов (подмножеств для $\boldsymbol{\Omega}$), так и подмножеств образующих (подмножеств для $\boldsymbol{U}$).

Неподвижными точками замыкания Галуа $[\_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ будем называть все такие $[X]_{\boldsymbol{U},\boldsymbol{\Omega}}$, что:

$$[X]_{\boldsymbol{U},\boldsymbol{\Omega}} = X .$$

Простейшие представления о замыканиях Галуа и их неподвижных точках дает

Пример 1.

Положим $\boldsymbol{U} = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$, а $\boldsymbol{\Omega} = \{h_1, h_2, h_3, h_4\}$, где

$h_1 = \{a_1, a_2, a_4\}$,
$h_2 = \{a_1, a_2, a_5\}$,
$h_3 = \{a_1, a_3, a_6\}$,
$h_4 = \{a_1, a_3, a_7\}$.

Несложно убедиться, что:

$$[\{a_1\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1\},$$
$$[\{a_2\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1, a_2\} \supset^{6} \{a_2\},$$
$$[\{a_3\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1, a_3\} \supset \{a_3\},$$

тем не менее:

$h_1 = [\{a_1, a_2, a_4\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = [\{a_4\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1, a_2, a_4\} \supset \{a_4\}$,
$h_2 = [\{a_1, a_2, a_5\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = [\{a_5\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1, a_2, a_5\} \supset \{a_5\}$,
$h_3 = [\{a_1, a_3, a_6\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = [\{a_6\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1, a_3, a_5\} \supset \{a_6\}$,
$h_4 = [\{a_1, a_3, a_7\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = [\{a_7\}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \{a_1, a_3, a_7\} \supset \{a_7\}$.

□

Теперь обратимся к представлениям об «архитектуре» множества всех неподвижных точек определенного нами замыкания Галуа, порождаемых на заданном множестве исходных примеров (объектов) Ω. Рассмотрим следующий

Пример 2.

Пусть $\boldsymbol{U} = \{a_1, a_2, \dots, a_{20}\}$, $\boldsymbol{\Omega} = \{h_1, h_2, \dots, h_{13}, h_{14}\}$

и

$h_1 = \{a_1, a_2\}$, $h_2 = \{a_1, a_6, a_9\}$, $h_3 = \{a_2, a_3, a_4\}$,
$h_4 = \{a_2, a_3, a_5\}$, $h_5 = \{a_2, a_4, a_5\}$,
$h_6 = \{a_3, a_4, a_5\}$, $h_7 = \{a_4, a_5\}$, $h_8 = \{a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{16}\}$,
$h_9 = \{a_2, a_3, \dots, a_5, a_{10}\}$, $h_{10} = \{a_1, a_2, \dots, a_6, a_{14}, a_{15}\}$,
$h_{11} = \{a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{15}, a_{20}\}$,
$h_{12} = \{a_1, a_2, \dots, a_6, a_{12}, a_{17}\}$,
$h_{13} = \{a_1, a_2, \dots, a_6, a_{13}, a_{18}\}$, $h_{14} = \{a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{15}, a_{19}\}$.

Диаграмма $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ взаимной вложимости неподвижных точек замыкания Галуа для заданного множества объектов $\boldsymbol{\Omega}$ в данном случае имеет представленный на рис.1 вид.

По определению *архитектуры типа* α формируются всеми парами соседних вершин $<h_1,h_2>$ следующего вида:

(α_1) $\{a_1, a_2, \dots, a_k\} = h_1 \subset h_2 = \{a_1, a_2, \dots, a_k, a_{k+1}, \dots\}$,

(α_2) среди образующих – элементов множества h_2 – имеется такая a_0, что $h_2 = [a_0]$,

(α_3) для каждой вершины (множества образующих) h_1 множество h_2, формирующее вместе с ней рассматриваемую пару $<h_1,h_2>$, представляет собой минимальный элемент в $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$, удовлетворяющий условиям (α_1) и (α_2).

[3] Т.е. множество объектов, построенных из образующих $a_1, a_2, \dots, a_n$.

[4] См., например, [10,11] и др.

[5] См. предыдущую сноску.

[6] Т.е. множество $\{a_2\}$ есть подмножество множества $\{a_1, a_2\}$, не совпадающее с последним.

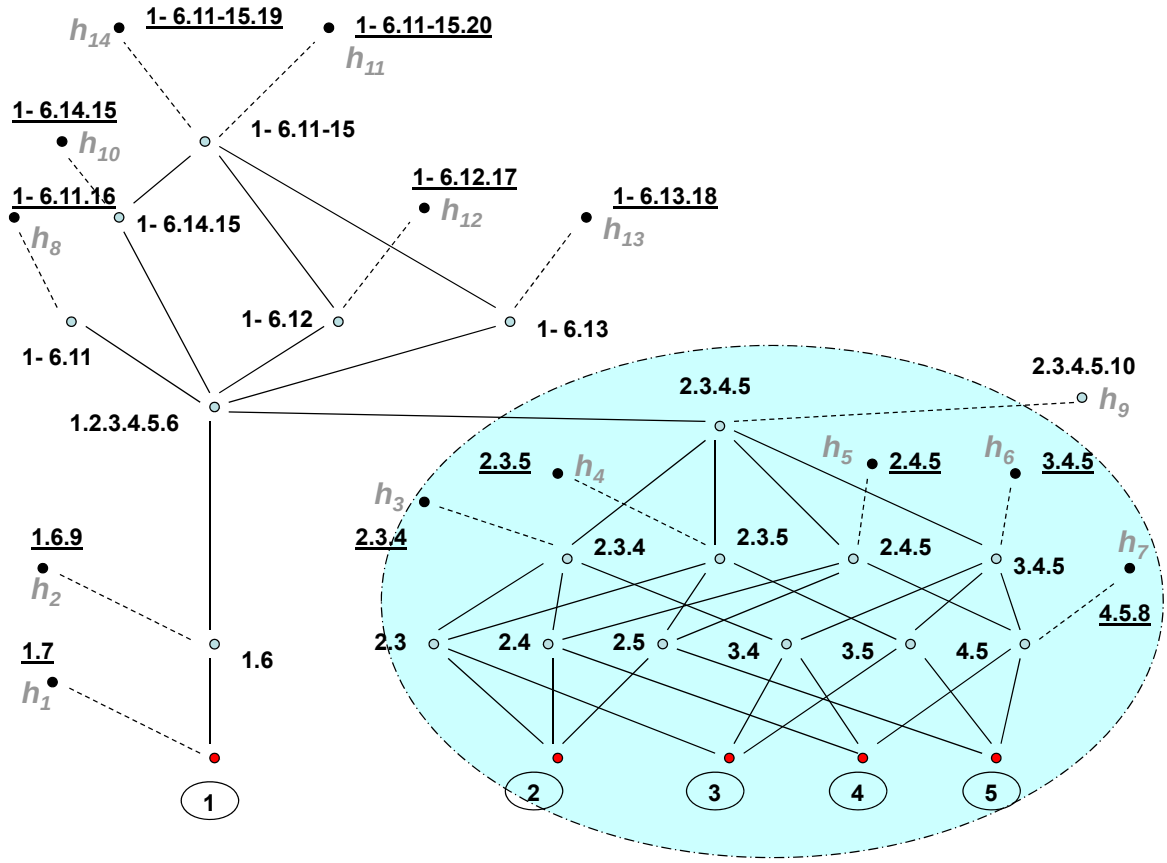

Рис. 1. Диаграмма $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ для заданного множества $\boldsymbol{\Omega}$

В свою очередь, *архитектуры типа* β – это комбинаторно сложные (т.е. «накрывающие» его существенную часть) фрагменты соответствующего гиперкуба, вкладывающегося в рассматриваемую диаграмму $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

С формальной точки зрения фрагменты типа β определяются следующим образом: они формируются всеми такими вершинами рассматриваемого псевдодерева[7], которые:

(β_1) составляют нижнюю границу соответствующего гиперкуба (т.е. при порождении замыканий одноэлементных подмножеств находящегося в корне рассматриваемого псевдодерева примера – множества образующих – и формировании из них каркаса этого псевдодерева, эти вершины оказываются наименьшими вершинами одного уровня в смысле упорядочения по вложению множеств образующих, сопоставленных вершинам порожденного каркаса), или же

(β_2) находятся на более «высоких» (по отношению к только что обсуждавшейся нижней границе гиперкуба) уровнях, при этом

(β_3) в множестве образующих, сопоставленных каждой такой вершине, нельзя найти элемент, замыкание которого совпадает с этим (сопоставленным данной вершине) множеством образующих.

□

Другими словами: фрагменты типа β формируются за счет «равнозначных» (одноуровневых однотипных) вершин каркаса и нижней границы каждого соответствующего гиперкуба, а также вершин, которым сопоставлены все допустимые комбинации множеств образующих, соответствующих только что выбранным на нижней границе вершинам каждого такого гиперкуба.

Наконец, *архитектуры типа* γ формируются «наложением» фрагментов типа α и\или β на некоторые подмножества веток в уже построенном ранее[8] фрагменте типа β. Фактически здесь мы имеем дело либо с «одиночными» цепями частичного порядка[9] типа α, или же с множествами таких цепей, представленными фрагментами дополнительных[10] булевских кубов типа β, причем и те и другие «наложены» на некоторые ветки «ранее» (т.е. – в смысле избранного порядка восстановления анализируемых диаграмм: от его листьев к корню) уже сформированных фрагментов

[7] Специального вида подграфа диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$, более подробное обсуждение свойств и особенностей которого будет представлено ниже (перед ***Примером 4***).

[8] При формировании соответствующего псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{A\}}$ *от* его *листьев* к *корню*.

[9] В смысле взаимной вложимости подмножеств образующих из алфавита $\boldsymbol{U}$.

[10] Полученные за счет выявления «новых» «склеенных» образующих.

типа β в рассматриваемой нами исходной диаграмме $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

Таким образом, цепочки частичного порядка, образованные (см. рис. 1), например, вершинами

$$< \{ a_1\}, \{ a_1, a_6\}, \{ a_1, a_2, \dots , a_6 \}, \{ a_1, a_2, \dots , a_6 , a_{11} \}, \{ a_1, a_2, \dots , a_6 , a_{11} , a_{16}\} >$$

или

$$< \{ a_1\}, \{ a_1, a_6\}, \{ a_1, a_2, \dots , a_6 \}, \{ a_1, a_2, \dots , a_6 , a_{12} \}, \{ a_1, a_2, \dots , a_6 , a_{11} , a_{12} , \dots , a_{15}\}, \{ a_1, a_2, \dots , a_6 , a_{11} , a_{12} , \dots , a_{15}, a_{20} \} >,$$

являются примерами архитектур типа α.

В свою очередь, выделенная темным фоном в овале на рис.1 область (исключая вершины h_3, h_4, … ,h_7) – пример архитектуры типа β.

Наглядные представления об архитектурах типа γ помогает сформировать

Пример 3.

Пусть $\boldsymbol{U} = \{a_1, a_2, \dots , a_9\}$, $\boldsymbol{\Omega} = \{ h_1, h_2, \dots , h_7\}$

и

$h_1 = \{ a_1, a_2, a_3, a_5\}$, $h_2 = \{ a_1, a_2, a_4 \}$, $h_3 = \{ a_1, a_3, a_4 \}$,
$h_4 = \{ a_1, a_2, \dots , a_5 , a_8 \}$,
$h_5 = \{ a_1, a_2, \dots , a_5 , a_9 \}$, $h_6 = \{ a_2, a_3, \dots, a_6 \}$,
$h_7 = \{ a_1, a_2, \dots , a_5 , a_7 \}$.

Диаграмма $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ в данном случае имеет представленный на рис.2 вид.

Здесь цепочки частичного порядка, образованные, например, следующими вершинами (см. выделения жирным на рис.2)

$$< \{ a_2, a_3, a_5 \}, \{ a_2, a_3, a_4 , a_5\}, \{ a_1, a_2, \dots , a_5\}, \{ a_1, a_2, \dots , a_5 , a_9 \} >$$

или

$$< \{ a_2, a_3, a_5\}, \{ a_2, a_3, \dots , a_5\}, \{ a_1, a_2, \dots , a_5 , a_7\} >,$$

являются примерами архитектур типа γ.

Параллельно, выделенная темным фоном в овале на рис.2 область (исключая вершины h_1, h_2, … , h_7), – еще один пример архитектуры типа β.

Одним из промежуточных элементов обсуждаемого подхода является техника разбиения рассматриваемой диаграммы взаимного вложения неподвижных точек замыкания Галуа на «самостоятельные» части (поддиаграммы) таким образом, что

- объединение всех подобных «самостоятельных» частей совпадало бы с исходной диаграммой;
- каждая из таких самостоятельных частей порождалась бы «достаточно просто»;

и, наконец,

- число таких «самостоятельных частей» с ростом обучающей выборки и\или исходного алфавита росло бы «не очень быстро».

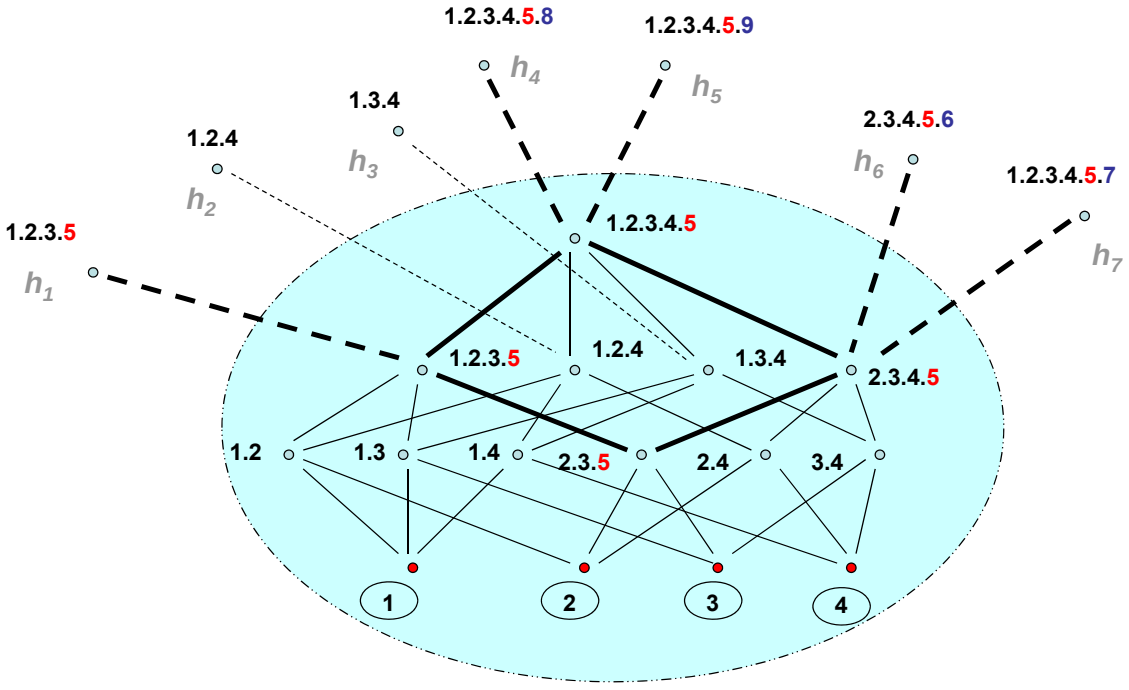

Рис. 2. Пример архитектуры типа γ

Такая схема реализуется путем выделения в исходной диаграмме отдельных поддиаграмм[11] (каждая из которых в корне имеет один из примеров, входящих в исходную обучающую выборку, а ее ветки ведут к каждой из вершин, образованной минимальными по вложению ДСМ-сходствами – всем однобуквенным[12] неподвижным точкам рассматриваемого замыкания Галуа для букв исходного алфавита, входящих в выбранный в качестве корневой вершины этой поддиаграммы объект – пример из обучающей выборки). Подобные поддиаграммы мы будем называть псевдодеревьями, подчеркивая, что поддиаграммы предлагаемого вида, вообще говоря, могут и не быть деревьями (например, за счет существования в них пар вершин во фрагментах типа β, находящихся, по крайней мере, на двух различных ветках частичного порядка взаимной вложимости соответствующих замыканий Галуа), однако, они могут быть сведены к виду дерева удалением тех или иных подмножеств входящих в них ребер. (В свою очередь, каждое такое псевдо-дерево может быть представлено как объединение соответствующих деревьев, содержащих то же самое множество вершин и вкладывающихся в рассматриваемое псевдодерево).

Особенности архитектуры диаграмм и входящих в них псевдодеревьев нам поможет проиллюстрировать

Пример 4.

Пусть

$\mathbf{U} = \{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_{28}}\}$
$\mathbf{\Omega} = \{ h_1, h_2, \ldots, h_{18} \}$

h_1: **1.2.3.5.10.11**
h_2: **1.2.4.10.11.13**
h_3: **1.3.4.10.11.13**
h_4: **1.2.3.4.5.9.10.11.13.26**
h_5: **2.3.4.5.6.10.11.13.26.27**
h_6: **2.3.4.5.7.10.11.13.26**
h_7: ***10.11***
h_8: **11.12**
h_9: **14.15**
h_{10}: **14.16.17**
h_{11}: *1.2.3.4.5.8.10.11.13.***14.16.18.19.26**
h_{12}: *1.2.3.4.5.8.10.11.13.***14.16.18.20.26**
h_{13}: *1.2.3.4.5.8.10.11.13.***14.16.22.26**
h_{14}: *1.2.3.4.5.8.10.11.13.***14.16.21.26**
h_{15}:*1.2.3.4.5.8.10. 11.***13.14.16.23.26**
h_{16}: *1.2.3.4.5.8.10.11.13.***14.16.21.22.23.24.26**
h_{17}: *1.2.3.4.5.8.10.11.13.***14.16.21.22.23.25.26**
h_{18}: **2.3.4.5.6.10.11.13.26.28**

В данном случае диаграмма $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ будет иметь следующий вид (рис. 3).

При классификации вершин рассматриваемых диаграмм следует учесть также и случай, когда в тех или иных вершинах типа β, расположенных на нижней границе соответствующего гиперкуба, появляются одновременно по две (или более) «равноправных» образующих. Фактически, это случай – будем называть его эффектом «склеенных» образующих (см., например, вхождения образующих $\mathbf{a_4}$ и $\mathbf{a_{13}}$ в псевдодерево на рис. 4), – когда каждую такую пару (или, соответственно более) образующих можно заменить одной вновь введенной в рассмотрение образующей, и при этом получить псевдодерево той же конфигурации[13].

Полезную информацию о внутренней структурной организации каждого из рассматриваемых псевдодеревьев может предоставить его специальный подграф – каркас. По определению *каркасом* псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{A\}}$ мы будем называть граф

$$GK_{\mathbf{D,A}} = <\mathbf{K_{D,A}}, \mathbf{R_{D,A}}>,$$

множество $\mathbf{K_{D,A}}$ вершин которого образовано замыканиями всех образующих из выбранного множества (примера) $A = \{a_1, a_2, \ldots, a_n\}$:

$$\mathbf{K_{D,A}} = \{[a_1]_{U,\Omega}, [a_2]_{U,\Omega}, \ldots, [a_n]_{U,\Omega}\},$$

а множество ребер $\mathbf{R_{D,A}}$ отражает «ближайшие»[14] попарные взаимные вложимости элементов из $\mathbf{K_{D,A}}$ друг в друга.

Так, например, для представленного на рис. 4 псевдодерева с корнем в вершине

h_{17}:*1.2.3.4.5.8.10.11.13.***14.16.21.22.23.25.26**

его каркас $GK_{\mathbf{D},h_{17}}$ формируется следующими замыканиями:

[a_1] : 1.10.11
[a_2] :2.10.11
[a_3] :3.10.11
[a_4] :4.10.11.13
[a_5] :2.3.5.10.11
[a_8] :*1.2.3.***4**.5.8.10.11.***13.14.16.26**
[a_{10}] :10.11
[a_{11}] :11
[a_{13}] :4.10.11.13
[a_{14}] :14
[a_{16}] :14.16
[a_{21}] :*1.2.3.4.5.8.10. 11.***13.14.16.21.26**
[a_{22}] :*1.2.3.***4**.5.8.10. 11.***13.14.16.22.26**
[a_{23}] :*1.2.3.4.5.8.10. 11.***13.14.16.23.26**
[a_{25}] :*1.2.3.4.5.8.10.11.13.***14.16.21.22.23.25.26**
[a_{26}] :2.3.4.5.10.11.13.26

собранными в граф следующего вида (см. рис 5).

[11] Подграфов специального вида.

[12] Вообще говоря, здесь в качестве листьев рассматриваемой поддиаграммы могут встречаться и неоднобуквенные подмножества исходно заданного алфавита – так называемые «склеенные образующие» (исключительно вместе встречающиеся и во всех остальных вершинах этой поддиаграммы). Важно лишь заметить, что каждый такой лист есть минимальное (по вложению соответствующих подмножеств образующих) множество из сопоставленных вершинам рассматриваемой поддиаграммы букв исходно заданного алфавита. Можно показать, что существует быстрый алгоритм диагностики «склеенности» образующих в соответствующей поддиаграмме.

[13] Т.е. псевдодерево с тем же множеством ребер и с тем же (пусть и сформированным заменой «склеенных» образующих вновь введенной образующей) множеством вершин.

[14] Т.е. если $\{\mathbf{a},\mathbf{b}\} \subseteq \mathbf{K_{D,A}}$, то ребро $<\mathbf{a},\mathbf{b}>$ принадлежит $\mathbf{R_{D,A}}$тогда и только тогда, когда $\mathbf{a} \subseteq \mathbf{b}$, и неверно, что в $\mathbf{K_{D,A}}$ найдется (отличный от **b**) элемент **c,** такой, что $\mathbf{a} \subseteq \mathbf{c} \subseteq \mathbf{b}$.

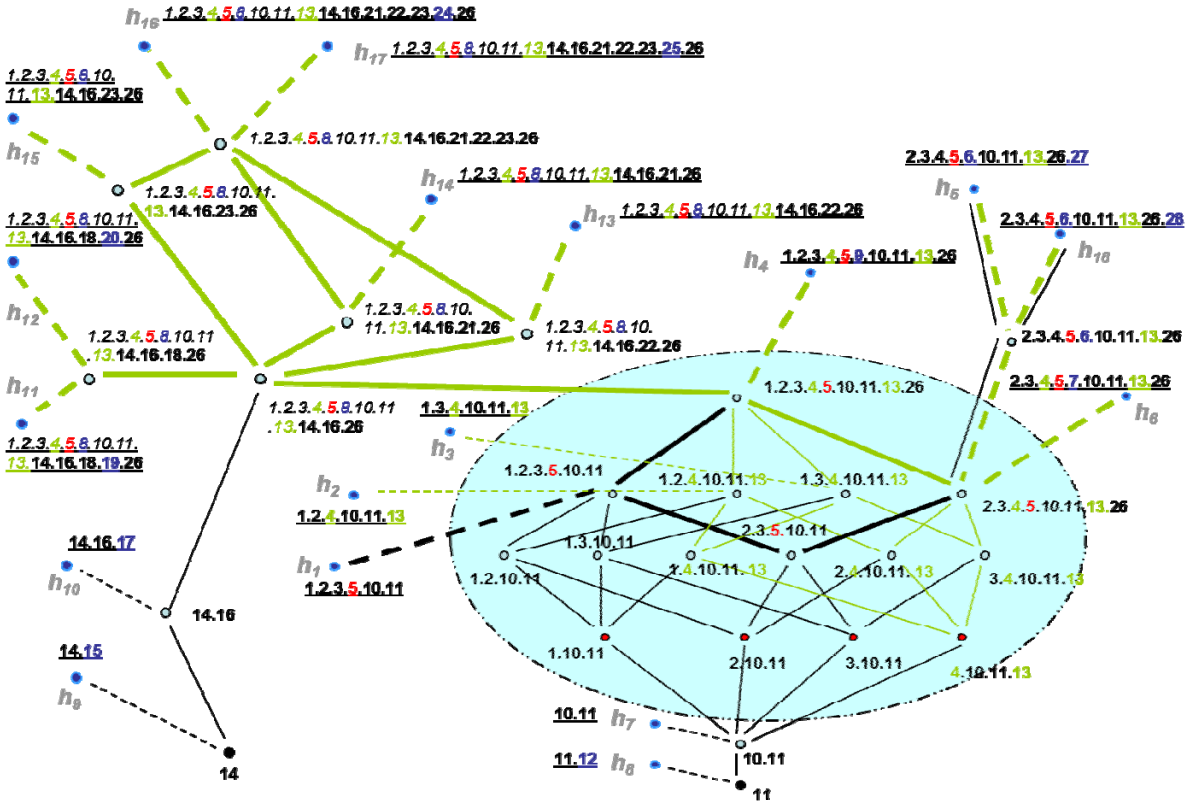

Рис. 3. Диаграмма $\mathbf{D\_GC}_{\varphi,\Gamma}(\boldsymbol{U},\boldsymbol{\Omega})$ для рассматриваемого случая

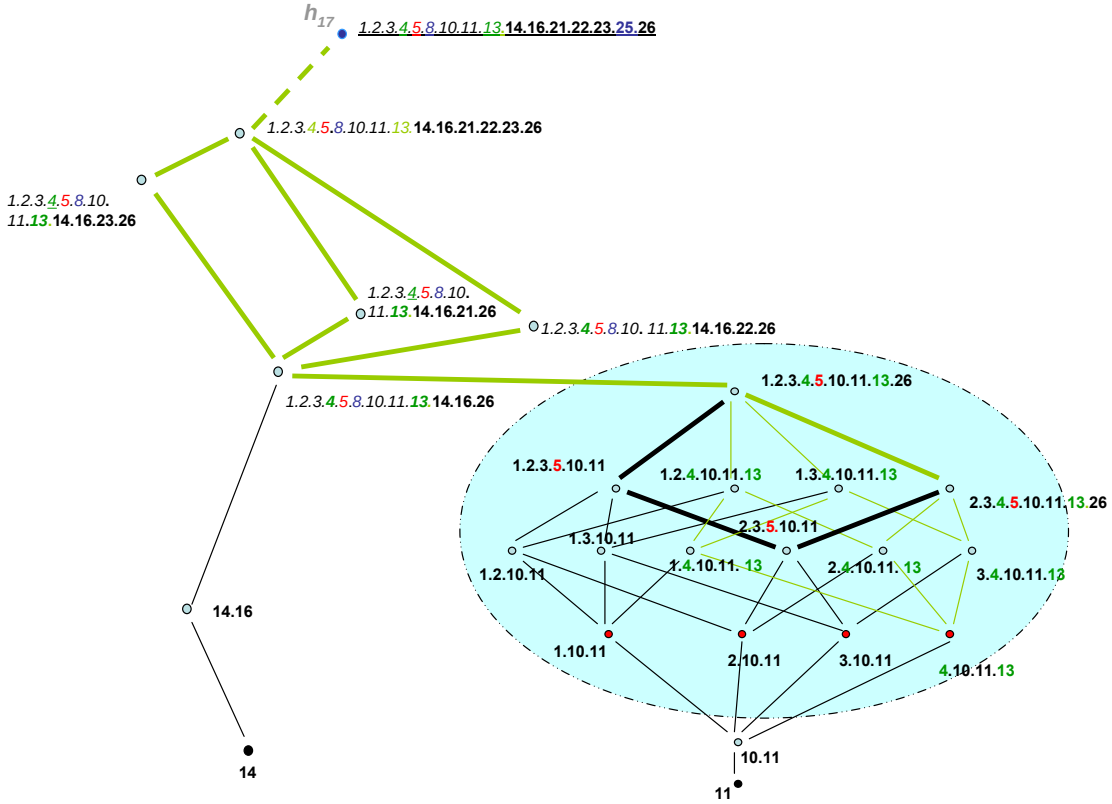

Рис. 4. Примеры встречаемости «склеенных» и «наклеенных» образующих:
a) «склеенные» образующие $\mathbf{a}_4$ и $\mathbf{a}_{13}$ – вершина <*4*.10.11.*13*>,
b) «наклеенная» (на образующие $\mathbf{a}_2$, $\mathbf{a}_3$, $\mathbf{a}_{10}$ и $\mathbf{a}_{11}$) образующая $\mathbf{a}_5$ – вершина <2.3.***5***.10.11>.

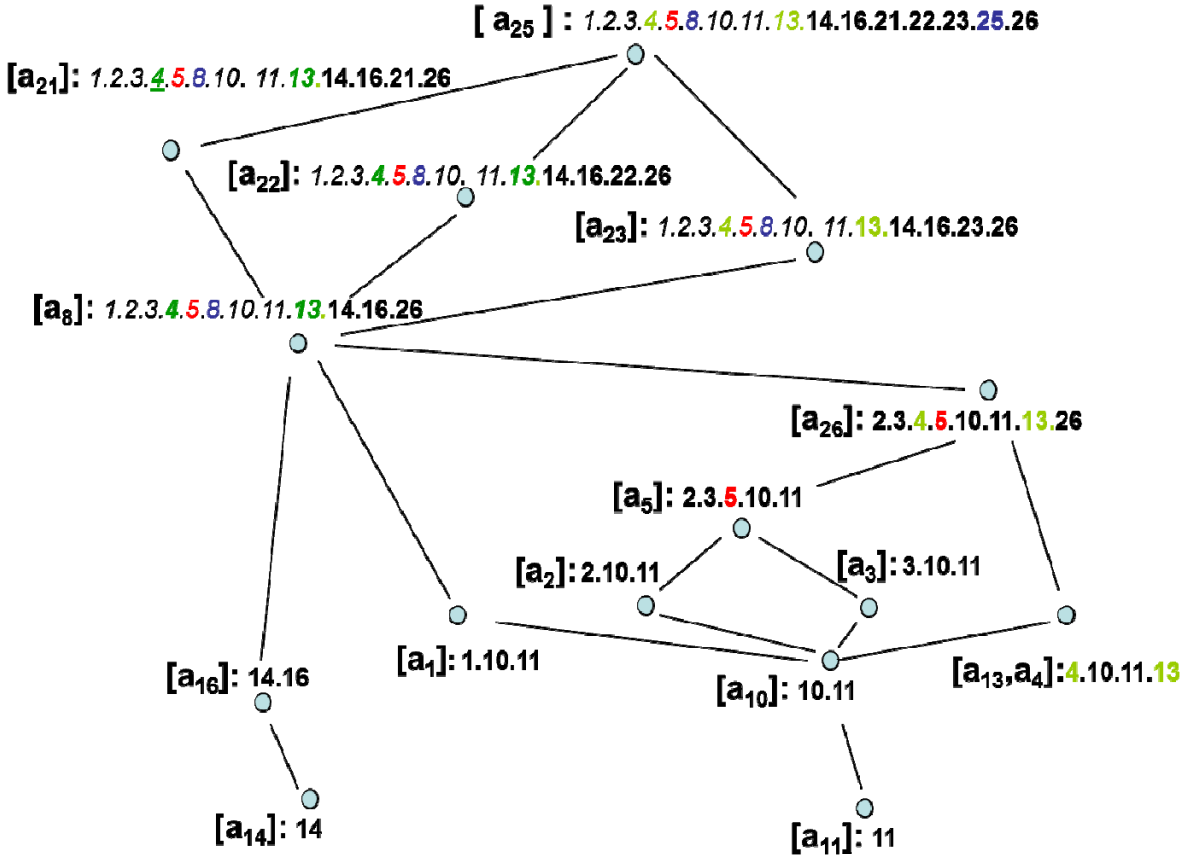

Рис. 5. Каркас $GK_{\mathbf{D},h17}$ псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|\,\boldsymbol{h_{17}}$:

Для заданного псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$ процедура формирования каркаса $GK_{\mathbf{D},(\boldsymbol{\Omega}|\{A\})}$ выглядит следующим образом:

- формируем замыкания Галуа $[\_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ для всех элементов текущего алфавита $\boldsymbol{U}$;

- в полученном множестве замыканий всех однобуквенных подмножеств $\boldsymbol{U}$ сначала выделяем все минимальные элементы, после этого (на каждом шаге контролируя условия «диагностики» формирования в анализируемом $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$ архитектур типа β, а с каждым из таких случаев – и булевских гиперкубов соответствующей размерности);

- «поэтажно» восстанавливаем каркас $GK_{\mathbf{D},(\boldsymbol{\Omega}|\{A\})}$, последовательно выделяя «ближайшие» по вложению элементы построенного множества замыканий Галуа $[\_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ всех однобуквенных подмножества алфавита $\boldsymbol{U}$.

Для представленного на рис. 5 случая эта процедура будет выглядеть следующим образом:

- на первом шаге формируем множество замыканий для всех однобуквенных подмножеств алфавита $\boldsymbol{U}$ (см. их перечень выше, перед рис. 5). Далее

- выделим минимальные элементы (это будут $\mathbf{[a_{11}]} =: \{\mathbf{a_{11}}\}$ и $\mathbf{[a_{14}]} = \{\mathbf{a_{14}}\}$). Следующим шагом находим для каждого из них ближайшие надмножества: для $\mathbf{[a_{11}]}$ это – $\mathbf{[a_{10}]}=\{\mathbf{a_{10}},\mathbf{a_{11}}\}$, а для $\mathbf{[a_{14}]}$ – это $\mathbf{[a_{16}]}=\{\mathbf{a_{14}},\mathbf{a_{16}}\}$. Затем таким же способом переходим к следующему «этажу» каркаса $GK_{\mathbf{D},h17}$ – вершине $\mathbf{[a_8]}$ и четверке вершин $\{\mathbf{[a_1]}, \mathbf{[a_2]}, \mathbf{[a_3]}, \mathbf{[a_4]}\}$. При этом проверка условий появления архитектур типа β показывает, что эта четверка – нижняя граница (см. подробнее [8, 9]) четырехмерного гиперкуба[15] $\mathbf{B_4}$, некоторый фрагмент которого должен будет войти в псевдодерево $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|\,\boldsymbol{h_{17}}$. (Таким образом, при восстановлении всего псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|\,\boldsymbol{h_{17}}$ по каркасу $GK_{\mathbf{D},h17}$ нам придется далее заняться «реконструкцией» в том числе и той части анализируемого псевдодерева, которая «наложена» на гиперкуб $\mathbf{B_4}$). Далее, переходя к следующему «этажу» каркаса $GK_{\mathbf{D},h17}$, выделяем вершины $\mathbf{[a_5]}$, затем $\mathbf{[a_{26}]}$, после чего – $\mathbf{[a_8]}$[16]. Продолжая далее представленный процесс восстановлением каркаса $GK_{\mathbf{D},h17}$ «этаж» за «этажом», финишируем в вершине $\mathbf{[a_{25}]}$, соответствующей объекту

$$\boldsymbol{h_{17}}:\mathit{1.2.3.4.5.8.10.11.13}\mathbf{.14.16.21.22.23.25.26}$$

из исходного множества $\boldsymbol{\Omega}$.

[15] Минимальный элемент которого - $\mathbf{[a_{10}]} = \{\mathbf{a_{10}},\mathbf{a_{11}}\}$, а максимальный представляет собой соответствующее подмножество для $\mathbf{[a_8]}$.

[16] С учетом также выявленной «*кросс-этажной*» вложимости $\mathbf{[a_{16}]}$ в $\mathbf{[a_8]}$.

При восстановлении каждого «наложенного» на соответствующий ему гиперкуб $\mathbf{B_s}$ фрагмента анализируемого псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$ необходимо обратить внимание на следующие три обстоятельства:

- необходимо идентифицировать множество $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$ всех «раскрывающихся» (см. подробнее [8, 9]) в этот гиперкуб вершин диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$, представляющее собою определенное подмножество исходно заданного множества примеров $\boldsymbol{\Omega}$. При этом $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$ формируется из множества $\boldsymbol{\Omega}_{\mathbf{Bs}}$ раскрывающихся собственно в гиперкуб $\mathbf{B_s}$ примеров из $\boldsymbol{\Omega}$, пополненного теми уже восстановленными при порождении каркаса $GK_{\mathbf{D},(\boldsymbol{\Omega}|\{A\})}$ ДСМ-сходствами (псевдопримерами[17]), которые «представляют»[18] оставшиеся примеры из $\boldsymbol{\Omega}$, в свою очередь раскрывающиеся в гиперкуб $\mathbf{B_s}$ через соответствующие вершины псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$, а затем – и теми сходствами из $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$, которые раскрываются в $\mathbf{B_s}$ через вершины диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$, не вошедшие[19] в диаграмму $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$. (При этом в данном случае, как и в случае рассмотренного чуть выше непрямого раскрытия примеров из $\boldsymbol{\Omega}$ через соответствующие вершины диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$, формируемое дополнительное пополнение для множества $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$ строится с учетом минимального представительства выбираемыми ДСМ-сходствами соответствующих примеров из исходного $\boldsymbol{\Omega}$);

- далее следует рассматривать $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$ как новое целевое множество примеров для анализа порождаемых с его помощью множества неподвижных точек используемого замыкания Галуа $[\_]_{\boldsymbol{U}(\mathbf{Bs}),\boldsymbol{\Omega}(\mathbf{Bs})}$ и диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega}^*_{\mathbf{Bs}})$, т.е. выполнить шаг рекурсивной процедуры от ДСМ-анализа исходного множества примеров $\boldsymbol{\Omega}$ к восстановлению ДСМ-сходств из его подмножества $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$. (При этом можно показать, что за исключением специальных быстро диагностируемых случаев[20] данная рекурсивная процедура порождает сжимающее отображение

$$\boldsymbol{\Omega} \rightarrow \boldsymbol{\Omega}^*_{\mathbf{Bs}},$$

т.е. позволяет за конечное[21] число шагов проанализировать все исходно заданное множество примеров $\boldsymbol{\Omega}$);

и, наконец,

- выбор анализируемых в первую очередь псевдодеревьев для «нового» (см. только что представленную рекурсивную процедуру) множества примеров $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$ может быть выполнен целенаправленно – с учетом тех или иных «управляющих» условий (например, с учетом структуры[22] представленных для ДСМ-прогноза новых объектов – примеров из соответствующего множества $\boldsymbol{\Omega}_{\boldsymbol{\tau}}$). Таким образом, можно организовать целенаправленную «навигацию» в множестве всех ДСМ-сходств в процессе восстановления целевой диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ для исходно заданного множества примеров $\boldsymbol{\Omega}$, порождая в первую очередь «полезные» (например, для ДСМ-прогноза) эмпирические зависимости, а уж затем, если потребуется, и все оставшиеся элементы диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$. Следовательно, появляется возможность организовать *целенаправленно управляемые приближенные вычисления* при порождении эмпирических зависимостей в рамках ДСМ-метода. При этом предельным «приближением» в рамках предлагаемой процедуры будет порождение собственно всех ДСМ-сходств, формируемых из заданного множества примеров $\boldsymbol{\Omega}$, а с ними – и всей диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

Рассмотрим, как обсуждаемый переход от $\boldsymbol{\Omega}$ к $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$ («шаг» представленной рекурсии) выглядит на примере. Выявленный в ходе восстановления каркаса $GK_{\mathbf{D},h17}$ гиперкуб $\mathbf{B_4}$ на рис. 3-4 выделен пунктирным овалом. Раскрывающиеся в него примеры исходного множества $\boldsymbol{\Omega}$ (терминальные вершины псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{h17}$, дополненные соответствующими «представителями» тех терминальных вершин, которые раскрываются не прямо в $\mathbf{B_4}$, а лишь через некоторые «внутренние» элементы соответствующего множества всех ДСМ-сходств) – это:

- $\boldsymbol{h_1}$, $\boldsymbol{h_2}$,$\boldsymbol{h_3}$, $\boldsymbol{h_4}$, $\boldsymbol{h_6}$ (как раскрывающиеся непосредственно в гиперкуб $\mathbf{B_4}$ примеры из $\boldsymbol{\Omega}$), а также – «представители» остальных релевантных рассматриваемому архитектурному фрагменту типа β примеров из $\boldsymbol{\Omega}$:

- $\boldsymbol{h_5}$ и $\boldsymbol{h_{18}}$ (через не входящую в псевдодерево $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{h17}$ вершину $\mathbf{[a_6]}=\{\mathbf{a_2},\mathbf{a_3},\mathbf{a_4},\mathbf{a_5},\mathbf{a_6},\mathbf{a_{10}},\mathbf{a_{13}},\mathbf{a_{26}}\}$ сводной диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ всех ДСМ-сходств),

- $\boldsymbol{h_{11}}$–$\boldsymbol{h_{17}}$ (через вершину $\mathbf{[a_8]}=\{\mathbf{a_1},\mathbf{a_2},\mathbf{a_3},\mathbf{a_4},\mathbf{a_5},\mathbf{a_8},\mathbf{a_{10}},\mathbf{a_{11}},\mathbf{a_{13}},\mathbf{a_{14}},\mathbf{a_{16}},\mathbf{a_{26}}\}$ диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{h17}$),

таким образом, в расширенное множество $\boldsymbol{\Omega}^*_{\mathbf{B4}}$ мы поместим

- входящие в исходное множество $\boldsymbol{\Omega}$ примеры $\boldsymbol{h_1}$, $\boldsymbol{h_2}$, $\boldsymbol{h_3}$, $\boldsymbol{h_4}$, $\boldsymbol{h_6}$, которые и сформируют собственно множество $\boldsymbol{\Omega}_{\mathbf{B4}}$, затем дополнив его

- «представляющим» примеры $\boldsymbol{h_{11}}$–$\boldsymbol{h_{17}}$ из $\boldsymbol{\Omega}$ псевдопримером $\mathbf{[a_8]}$,

- «представляющим» примеры $\boldsymbol{h_5}$ и $\boldsymbol{h_{18}}$ из $\boldsymbol{\Omega}$ псевдопримером $\mathbf{[a_6]}$, который будет (полиномиально быстро по отношению к размерам множеств $\boldsymbol{U}$ и $\boldsymbol{\Omega}$) построен в два этапа: сперва перебором по $\boldsymbol{\Omega}$ будут выделены примеры $\boldsymbol{h_5}$ и $\boldsymbol{h_{18}}$, а затем взятием замыканий, входящих в них одноэлементных подмножеств

---

[17] Т.е. примерами лишь в смысле нового множества $\boldsymbol{\Omega}^*_{\mathbf{Bs}}$. В исходном множестве $\boldsymbol{\Omega}$ их нет.

[18] Как минимальные надмножества лежащих на гиперкубе $\mathbf{B_s}$ элементов каркаса $GK_{\mathbf{D},(\boldsymbol{\Omega}|\{A\})}$, множество которых содержит для каждого примера из исходного $\boldsymbol{\Omega}$ минимальное число «представляющих» его псевдопримеров (по крайней мере один).

[19] Это обстоятельство может быть диагностировано с учетом структуры подмножества элементов исходного алфавита $\boldsymbol{U}$, задействованных при формировании псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\,|_{\{A\}}$ (т.е. входящих в корневую вершину этой диаграммы).

[20] Ситуаций, когда $\boldsymbol{\Omega}=\{\boldsymbol{h_1},\boldsymbol{h_2},\ldots,\boldsymbol{h_s}\}$ – верхняя граница соответствующего гиперкуба $\mathbf{B_s}$ (возможно дополненная его наибольшим элементом $\boldsymbol{h_{s+1}}$).

[21] Более точно – ограниченное сверху числом элементов исходного множества примеров $\boldsymbol{\Omega}$.

[22] Входящих в них образующих из алфавита $\boldsymbol{U}$ и др.

алфавита ***U,*** будет выделен необходимый нам представитель – **[a_6]**.

Нетрудно убедиться, что

$$|\mathbf{\Omega}| > |\mathbf{\Omega}^*_{\mathbf{B4}}| \geq |\mathbf{\Omega}_{\mathbf{B4}}|,$$

т.е., выполняя «шаг» описанной рекурсивной процедуры, мы действительно имеем дело со сжимающим отображением.

Продолжим предпринятый нами процесс восстановления псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\,|\,\boldsymbol{h_{17}}$. В рамках выполняемого «шага» рекурсии проверим (см. подробнее [8, 9]) выполнимость критерия совпадения структуры гиперкуба $\mathbf{B_4}$ с накладываемым на него фрагментом диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\,|\,\boldsymbol{h_{17}}$. Так, принимая во внимание наличие в множестве $\mathbf{\Omega}^*_{\mathbf{B4}}$ объектов (примеров и псевдопримеров) ***h_1, h_2, h_3, h_4, h_6***, **[a_6]** и **[a_8]**, несложно убедиться, что с их помощью порождаются все четыре элемента верхней границы[23] гиперкуба $\mathbf{B_4}$. Следовательно, в архитектуре псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\,|\,\boldsymbol{h_{17}}$ будет «воспроизведен» весь рассматриваемый гиперкуб $\mathbf{B_4}$. Остается лишь уточнить (проверив замыкания одноэлементных подмножеств для ***h_1, h_2, h_3, h_4, h_6*** , **[a_6]** и **[a_8]**) явный вид расположенных в вершинах этого фрагмента псевдодерева $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\,|\,\boldsymbol{h_{17}}$ множеств образующих – элементов множества ***U***. Эта (полиномиально сложная – !) процедура и даст нам, как результат, финальный вид восстанавливаемого фрагмента, а с ним – и всей диаграммы $\mathbf{D\_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\,|\,\boldsymbol{h_{17}}$.

При порождении «полноформатных» ДСМ-зависимостей вида

< *объект*$\Rightarrow_1$*множество свойств* > вида (C$\Rightarrow_1$A)

а также

<*подобъект* (как причина, «носитель»)$\Rightarrow_1$*множество свойств*>вида (V$\Rightarrow_2$W)

(см. подробнее [1, 2] и др.) уже рассмотренная нами техника анализа примеров (элементов множества $\mathbf{\Omega}$) и их составных частей (подмножеств образующих алфавита ***U***) абсолютно аналогичным образом расширяется как на работу с контрпримерами, так и на операции с множествами свойств в правых частях формул с предикатами $\Rightarrow_1$ и $\Rightarrow_2$. При этом существенным оказывается порядок обработки левых и правых частей соответствующих формул: в подавляющем большинстве известных приложений с комбинаторной точки зрения правые части названных формул («отвечающие» за свойства изучаемых объектов а также их каузально-значимых подобъектов) устроены существенно проще, чем левые части (описывающие внутреннюю структуру этих объектов или же их подобъектов), Таким образом, старт процедур перебора с правых частей анализируемых формул с последующим применением аналогичных процедур уже к усеченным структурным описаниям в левых частях, как правило, предоставляет дополнительные возможности для оптимизации реализуемого ДСМ-перебора вариантов.

Аналогичным образом, при проверке дополнительных логических условий (запрета на контрпримеры, единственности причины и др.) при проверке выполнимости ДСМ-правил правдоподобного вывода первого рода (ППВ-I – см. подробнее [1, 2] и др.) запуск сначала проверки дополнительных условий (например, анализ так называемых *стоп-листов* при проверке запрета на контрпримеры – см. подробнее [8, 9]), а уж только «вторым темпом» – собственно ДСМ-анализа структур в левых частях соответствующих описаний примеров и контрпримеров, также несет в себе определенный потенциал сокращения ДСМ-перебора.

Наконец, при анализе выполнимости подходящих ДСМ-правил правдоподобного вывода второго рода (ППВ-II– см. подробнее [1, 2] и др.) также оказывается полезным вначале провести необходимые процедуры на правых частях описаний, содержащих предикаты $\Rightarrow_1$ и $\Rightarrow_2$ (например, проверить в рамках реализации ППВ-II наличие для доопределяемого множества свойств соответствующих покрытий множествами свойств из порожденных средствами ППВ-I эмпирических зависимостей – см. подробнее [1, 2, 8, 9] и др.), а уж только потом переходить к детальному анализу левых частей подходящих (характеризуемых подходящими множествами свойств) причинных зависимостей.

Весь комплекс описанных дополнительных процедур вместе с уже рассмотренной выше на примерах техникой целенаправленного и управляемого ДСМ-перебора вариантов (в том числе – представления целевой диаграммы вложения ДСМ-сходств как объединения самостоятельно обрабатываемых[24] псевдодеревьев, быстрого порождения каркасов соответствующих псевдодеревьев, восстановления анализируемых псевдодеревьев по их каркасам с использованием рекурсивных процедур восстановления архитектуры их фрагментов, наложенных на выявляемые в ходе формирования каркасов гиперкубы, и др.) формируют алгоритмический инструментарий ***приближенного*** ДСМ-метода. При этом имеются все необходимые возможности управления соответствующими ***приближениями***: от порождения сначала лишь (всех – !) полезных (например, для прогноза свойств конкретных новых объектов) ДСМ-зависимостей, далее к тем или иным управляемым расширениям этого множества зависимостей (например, к порождению всех эмпирических зависимостей, удовлетворяющих тем или иным заданным структурным ограничениям, – содержащим[25] те или иные наперед заданные комбинации образующих). И наконец, к

[23] Т.е. все четыре вершины этого гиперкуба $\mathbf{B_4}$, имеющие структуру соответствующего вида – <1,1,1,0>, <1,1,0,1>, <1,0,1,1><0,1,1,1>.

[24] В том числе (если это необходимо) – в параллельном режиме вычислений.

[25] В том числе – в виде заданных булевских комбинаций вида<*обязательная часть* (стабильное ядро) + ограниченно варьируемая *дополнительная часть* (дизъюнкция вхождений в структуру порождаемой эмпирической зависимости заданных множеств образующих) + произвольно варьируемая *«финальная» часть*>.

порождению[26] целиком всего множества ДСМ-зависимостей, формируемых из исходно заданного множества примеров и контрпримеров.

Наиболее привлекательной областью применения возможностей приближенного ДСМ-метода, по-видимому, является работа с большими выборками исходно заданных для порождения зависимостей примеров и контрпримеров. Однако в ряде практически значимых случаев (например, в реальных задачах выявления, анализа и устранения сбоев в сложных программно-технических комплексах – в области так называемых Hi-Tech-diagnostics&troubleshooting, о которых мы уже говорили ранее в настоящей работе, и др.) объем исходной выборки постоянно накапливаемых для последующего каузального анализа данных, не оставляет практически никаких реальных надежд на прямое использование и ДСМ-ИАД. В таких ситуациях, следуя высказанным еще Джоном Тьюки (John Wilder Tukey[27]) рекомендациям о целесообразности комбинирования статистических и дискретных «техник» восстановления «скрытых» в анализируемых выборках зависимостей, можно рассмотреть следующую методику ИАД для больших и сверхбольших[28] коллекций исходных данных:

ВХОД:

1) исходно заданная таблицей функция $\boldsymbol{F}$ (явный вид которой не известен):

значения (вектора) *переменных*	*значения функции* (на соответствующих векторах значений переменных)

При этом вектор $\mathbf{x}_0 = \langle x_1, x_2, \ldots, x_n \rangle$ в общем случае может содержать переменные двух типов, которые мы будем соответственно называть *каузальными* и *интерпретирующими*. Каузальные переменные (там, где можно «померять» их значения) отражают реально формирующие текущее значение функции $\boldsymbol{F}$ «влияния» (причинные факторы). Интерпретирующие переменные будут использованы для восстановления статистических зависимостей в данных (в таблице значений функции $\boldsymbol{F}$) – т.е. это некоторые факторы, «объясняющие» средствами соответствующей регрессии «поведение анализируемой функции $\boldsymbol{F}$ (причем – в рамках заданной нам таблицы значений – !)»;

2) набор заданных значений переменных (вектора) $\mathbf{x}_0$, для которого следует выполнить прогнозирование соответствующего значения изучаемой функции $\boldsymbol{F}$ ($\mathbf{x}_0$).

ВЫХОД:

1) прогнозируемое значение изучаемой функции $\boldsymbol{F}$ ($\mathbf{x}_0$);

2) система аргументов, позволяющая *на достаточном основании* принять результаты рассчитанного прогноза.

Процедура предложенного перехода **ВХОД-ВЫХОД** содержит следующие «шаги»:

- Для функции, исходно заданной таблицей значений (см. выше) средствами регрессий специального вида строится аналитическое приближение (аналитическая функция $\boldsymbol{F}^*(\mathbf{x}^*_0)$; от некоторого подмножества $\mathbf{x}^*_0$ переменных из набора $\mathbf{x}_0$[29].

- Параллельно для всех использованных в исходно заданной таблице векторов значений $\mathbf{x}_0$ уточняются текущие значения «каузальных» факторов – каузальных переменных, строится своего рода «Карта рисков» для анализируемых каузальных зависимостей.

- Сравнением поведения табличной функции $\boldsymbol{F}(\mathbf{x}_0)$ и восстановленной (как соответствующая регрессионная зависимость) функции $\boldsymbol{F}^*(\mathbf{x}^*_0)$ выделяются ситуации совпадения и несовпадения пиков (три случая[30]) табличного и аналитического представлений анализируемых данных.

- Карта рисков и представление «событий», описываемых исходной табличной функцией, множествами (и, возможно, отношениями на них) значений, характеризующими влияние рисков из Карты на конфигурацию каждой конкретной ситуации (т.е. окрестности соответствующего пика табличнозаданной функции) рассматривается как контекст для ДСМ-обучения в рамках ДСМ-ИАД.

- Прогнозирование новых значений табличной функции выполняется по следующей схеме: сначала на восстановленной (по исходной таблице) аналитической функции – соответствующей регрессии – вычисляется «базовое» значение для функции $\boldsymbol{F}$, которое далее корректируется за счет поправок, сформированных средствами ДСМ-обучения (в том числе – средствами приближенного ДСМ-метода) с учетом контекста влияния на конечный результат *определенных комбинаций* ранее выявленных факторов риска (значений каузальных переменных).

Эскиз применения подобной схемы интеллектуального анализа данных можно найти в работе [12], где предложенная А.А.Строевым техника восстановления статистических зависимостей в данных средствами так называемых панельных регрессий успешно использована при анализе поведения и прогнозировании поведения остатков на клиентских счетах двух известных (но существенно различающихся как по клиентской базе, так и по масштабам бизнеса) российских коммерческих банков. Далее – использование здесь соответствующей Карты рисков (реально влияющих на поведение остатков на клиентских счетах факторов – операций с остатками на счетах, приуроченных к календарным датам выплат

[26] «Оплаченному» исчерпывающим комбинаторным перебором и детальными расчетами соответствующей вычислительной сложности.

[27] См., например, http://ru.wikipedia.org/wiki/%D0%A2%D1%8C%D1%8E%D0%BA%D0%B8,_%D0%94%D0%B6%D0%BE%D0%BD и др.

[28] См. проблематику так называемых *bigdata*.

[29] Обычно это бывает некоторый «небольшой» набор интерпретирующих переменных.

[30] Т.е. – совпадение, несовпадение *вверх* (в соответствующей точке значение $\boldsymbol{F}^*(\mathbf{x}^*_0)$ выше значения $\boldsymbol{F}(\mathbf{x}_0)$) и несовпадение *вниз* (значение $\boldsymbol{F}^*(\mathbf{x}^*_0)$ ниже значения $\boldsymbol{F}(\mathbf{x}_0)$))

заработной платы сотрудникам, к штатным датам возвратных платежей по кредитам, к периодам налоговых выплат и т.п.) дает дополнительные возможности, позволяющие за счет учета результатов ДСМ-обучения на однородных группах событий (расположенных в окрестностях трех типов взаимного «соотнесения» пиков соответствующих табличной и аналитической функций[31]) рассчитать уточняющие поправки для соответствующего значения прогнозируемой функции (текущего объема остатков средств на счетах клиентов) на заданную текущую дату.

* * *

Итак, нами представлена техника обработки данных средствами ДСМ-ИАД, пригодная, в том числе, и для анализа больших выборок накапливаемых прецедентов (использования для организации машинного обучения и поддержки принятия управленческих решений). Приведенные нами аргументы дают основания надеяться, что предлагаемая технология

<*Псевдодеревья* + *Каркасы* + *Целенаправленная навигация*в множестве всех ДСМ-зависимостей>

позволит расширить границы применения ДСМ-ИАД и ДСМ-прогнозирования характеристик изучаемых процессов также и в промышленно-значимых приложениях. Дополнительные возможности здесь представляется естественным ожидать от комбинирования (например, предложенными выше способами) названных статистических подходов и ДСМ-ИАД.

СПИСОК ЛИТЕРАТУРЫ

1. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта // Искусственный интеллект и принятие решений. Часть I. – 2010. – № 3. – С.3 -21; Часть II. – 2010. – № 4. – С. 14-40.
2. Автоматическое порождение гипотез в интеллектуальных системах / ред. В.К.Финн. – М.: Либроком, 2009. – 528 С.

   3. Волкова А.Ю., Шестерникова О.П. О создании интеллектуальных систем, реализующих ДСМ–метод автоматического порождения гипотез, и результатах их применения для анализа медицинских данных // Научно-техническая информация. Сер. 2. – 2012. – № 5. – С. 10–15.
4. Волкова А.Ю. Опыт создания интеллектуальной ДСМ–системы для исследования данных различных предметных областей // Научно-техническая информация. Сер. 2. – 2013. – № 11. – С.12-26.
5. Михеенкова М.А., Волкова А.Ю. Спецификация интеллектуальной системы типа ДСМ // Научно-техническая информация**.** Сер. 2. – 2013. – № 7. – С. 5–19.
6. Шашкин Л.О. Применение методов эволюционного моделирования для оптимизации множества ДСМ-гипотез // Искусственный интеллект и принятие решений. – 2010. – № 3. – С. 33-39.
7. Шашкин Л.О. Сравнение приближенных способов поиска сходств для ДСМ-метода // Научно-техническая информация. Сер. 2. – 2010. – № 12. – С. 9-13.
8. Забежайло М.И. О некоторых возможностях управления перебором в ДСМ-методе // Искусственный интеллект и принятие решений. Часть I. – 2014. – № 1. – С.95-110.
9. Забежайло М.И. О некоторых возможностях управления перебором в ДСМ-методе // Искусственный интеллект и принятие решений. Часть II. – 2014. – № 3. – С.3-21.
10. Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод // Известия АН СССР. Сер. Техническая Кибернетика. –1987. – № 5. – С.42-63.
11. Кон П. Универсальная алгебра. – М.: Мир, 1968. – 359 С.
12. Строев А.А., Забежайло И.М. К проблеме прогнозирования объемов остатков средств на счетах коммерческих банков // Вестник Волгоградского Государственного Университета (в печати).

Материал поступил в редакцию 05.08.14.

Сведения об авторе

ЗАБЕЖАЙЛО Михаил Иванович – кандидат физико-математических наук, старший научный сотрудник, Управляющий директор НП «Центр прикладных исследований компьютерных сетей» (Москва, Сколково). e-mail: zmivan@gmail.com

[31]Т.е. для ДСМ-обучения здесь отдельным образом используются три группы примеров – обучение влиянию каузальных факторов при совпадении пиков, при несовпадении *вверх* и, отдельно, несовпадении *вниз*.

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 002 (051.2) – 047.44 : 001.891 – 047.72

К.И. Белоусов, Д.А. Баранов, Е.В. Ерофеева, Н.Л. Зелянская, Д.А. Ичкинеева

Прогнозирование научной области (на материале ведущего тематического журнала)*

Описывается исследовательская программа моделирования, оценки и прогнозирования состояния информационного пространства, сложившегося в отдельной предметной области, которая формируется ведущим научным журналом. В качестве предметной области рассмотрена когнитивная лингвистика, а материалом исследования послужили публикации российского научного журнала "Вопросы когнитивной лингвистики", целиком посвященного разработке лингвокогнитивной проблематики.

Ключевые слова: *научная предметная область, научный журнал, научная статья, термин, терминополе, графосемантическое моделирование, статистические методы, информационная система "Семограф"*

ВВЕДЕНИЕ

Научная предметная область (далее – ПрО) представляет собой создаваемую индивидуальными и коллективными агентами научного производства открытую мультиструктурную информационную систему, в которой осуществляется непрерывный интра- и интерпредметный процесс обмена информацией по доступным информационным каналам. Результатом информационных процессов становится постоянное обновление содержательной и структурной составляющих научной ПрО: конкуренция (изменение конфигураций, зависимостей, влияния друг на друга) частнонаучных ПрО в рамках общей предметной области, а также обновление концептуального пространства ПрО за счет междисциплинарных связей.

Научный журнал – одна из ключевых подсистем ПрО, являющаяся основным каналом обмена научной информацией, платформой для структурирования самой ПрО, а также средством

оценки эффективности научной деятельности и ценности ее результатов по отношению к агентам научного производства [1, 2]. Научные журналы становятся составной частью технологии выделения финансовой поддержки со стороны государства и частных научных фондов.

Несмотря на интенсификацию исследований в области моделирования ПрО науки (экономики, медицины и др.), научные журналы интересуют исследователей, как правило, в аспекте публикационной активности, а не того информационного поля, которое они формируют. Основу расчетов разнообразных индексов (цитируемости, самоцитируемости, Хирша, импакт-фактора и мн. др.) составляют сугубо формальные количественные показатели числа публикаций и цитирований за определенные периоды времени. Эти формальные показатели используются как основа планирования научной деятельности, в том числе в аналитических инструментах InCites и SciValSpotlight (приложениях, работающих на базе наукометрических показателей WoS и Scopus), созданных для оценки текущей деятельности агентов научного производства и применяемых как инструмент принятия решения в области финансирования проектов и научных коллективов [3–5].

Формальные инструменты дают показатели эффективности *результатов* научной деятельности, но не могут быть положены в основу планирования исследовательской деятельности в прогнозируемом состоянии ПрО, так как они не связаны с собственно информационным пространством ПрО. Этим обусловлена актуальность поисков новых форм, методов, инструментов моделирования информационного пространства, сложившегося в отдельной научной ПрО, для эффективного планирования научной деятельности индивидуальных и коллективных агентов научного производства. В этой связи представляют интерес модели информационного пространства ведущих научных журналов в исследуемой ПрО, так как ведущие научные журналы являются: а) главным "инструментом" разработки ПрО со стороны научного сообщества, б) средством ориентации исследователей, занимающихся разработкой отдельных аспектов ПрО, в самых последних разработках; в) средством диагностики и мониторинга состояния ПрО, г) конфигуратором частнонаучных ПрО (опре-

* Исследование выполнялось при финансовой поддержке РГНФ (проект № 12-34-01087) и РФФИ (проект РФФИ № 14-06-31143).

деляющим "вес" и положение отдельных направлений, школ, концепций относительно друг друга в рамках общей ПрО), д) проводником новых идей и направлений. Научный тематический журнал имеет не только синхронное измерение, но и диахронное, он представляет историю развития ПрО. Результаты исследования организации информационного пространства научного журнала (в том числе и в сопоставлении с организацией информационного пространства других научных журналов, в первую очередь зарубежных) могут быть полезны самым разным группам агентов научного производства.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Предлагаемая исследовательская программа изучения частнонаучных ПрО основывается на моделировании терминосферы ПрО. В качестве материала выступает корпус научных текстов, относящихся к ПрО.

В настоящей статье в качестве ПрО рассмотрена КОГНИТИВНАЯ ЛИНГВИСТИКА, а непосредственным материалом послужили публикации авторитетного российского научного журнала "Вопросы когнитивной лингвистики", целиком посвященного разработке лингвокогнитивной проблематики, с № 1 за 2008 г. по № 4 за 2013 г. (24 номера за 6 лет).

Операциональными единицами послужили базовые термины (в широком смысле слова), выносимые авторами публикаций в наборы ключевых слов (далее – НКС) к своим статьям. Обращение к ключевым словам (далее – КС) научных публикаций обусловлено тем, что, во-первых, посредством КС и НКС авторы сами обозначают доминанты концептуального пространства своих исследований; во-вторых, НКС научных публикаций представляют собой легко формализуемый конструкт в рамках большого корпуса текстов; в-третьих, НКС, в отличие от статьи, к которой они относятся, обычно доступны для автоматизированного извлечения.

В то же время необходимо остановиться на проблеме статуса КС и перспектив их использования в качестве операциональных единиц в нашем исследовании.

КС в научной статье представляют собой отдельные лексемы или словосочетания, *значительную часть которых нельзя отнести ни к терминосистеме, ни даже к терминологии изучаемой ПрО*. Одна (причем довольно объемная) часть КС к научным статьям представляет собой общенаучный слой терминологии (КАТЕГОРИЯ, СИСТЕМА, ПОНЯТИЕ и др.), другая содержит специальную лексику, характерную либо для языкознания в целом (ЯЗЫК, СЛОВО, ПРЕДЛОЖЕНИЕ и др.), либо для других его разделов/направлений (ДИСКУРС, ПОЛИТИЧЕСКАЯ КОММУНИКАЦИЯ, ИНТЕРТЕКСТУАЛЬНОСТЬ и др.). И, наконец, внушительная по объему часть КС относится не к терминологии, а к номенклатуре (ЛЕЗГИНСКИЙ ЯЗЫК, ЛЕКСЕМА "НОЧЬ", ОБРАЗ ЛОШАДИ и др.).

Таким образом, НКС к научным публикациям представляют собой предметно-понятийный субстрат (термины, номены, онимы), относящийся к разным подсистемам и структурам информационного пространства науки: терминологии и номенклатуре; общенаучным, дисциплинарным, частнонаучным (в том числе и смежным) ПрО. В НКС преломляются качества открытости, системности, полиструктурности, идущие одновременно диффузные и интеграционные процессы в самой науке.

Чтобы оценить соотношение терминологии ПрО (ее концептуального ядра) и сопутствующего номенклатурно-терминологического материала, достаточно сопоставить наши данные с известными корпусными исследованиями. Так, например, база данных ИНИОН по языкознанию [6], созданная на основе индексации 293440 документов, содержит в созданном на ее основе тезаурусе 1890 терминов (см. также [7]). В то же время материал нашего исследования включает 1650 КС к 446 научным публикациям.

Научная статья, репрезентирующая целостный фрагмент проведенного исследования, в полной мере обладает теми же качествами, что и "взрастившая" ее научная среда. Качества открытости, системности, диффузности и интегративности, полиструктурности проявляются в конструировании информационного пространства научной статьи и ее смысловых доминант, выносимых в композиционную область КС. Таким образом, термины, номены, онимы, не относящиеся напрямую к исследуемой ПрО, тем не менее крайне значимы для нее: *без этого предметно-понятийного субстрата невозможно представить ни синхронный срез ПрО, ни предложить ее ретроспективную или перспективную модель развития*.

Моделирование информационного пространства ПрО на основе множества понятий и реалий, репрезентированных в НКС, не в полной мере соответствует методам, используемым как в терминоведении, так и в онтологическом инжиниринге. В первую очередь речь идет о таком базовом методе исследования информационного пространства научной ПрО, как метод построения тезаурусов [8–11]. Тезаурус представляет собой модель языка науки, в которой связь между понятиями осуществляется "горизонтальными" и "вертикальными" семантическими типами отношений. В нашем случае сложность построения тезауруса обусловлена: 1) разнородностью материала, представляющего самые разнообразные стороны бытия науки, а не ПрО, которую он должен репрезентировать и структурировать; 2) неоднозначностью классификации – отнесения понятия к определенному узлу тезауруса. Сказанное в большой степени относится к терминологии и номенклатуре, так как термины и номены часто состоят из двух и более лексем, каждая из которых может быть отнесена к разным узлам тезауруса. Например, термин КОНЦЕПТУАЛЬНАЯ СТРУКТУРА ТЕКСТА был отнесен к полям КОНЦЕПТ И КОНЦЕПТОСФЕРА, (*концептуальная* структура текста), ТЕКСТ И ДИСКУРС (концептуальная структура *текста*) и СТРУКТУРНАЯ ОРГАНИЗАЦИЯ (концептуальная *структура* текста).

В то же время для нашего материала релевантен метод полевого анализа [12–14], с помощью которого структурируются семантические, понятийные, терминологические, тематические, ассоциативные и др. поля.

Остановимся на *принципах классификации КС.*

1. При отнесении термина (в широком смысле слова) к тому или иному терминополю рассматривалось наличие эксплицитной выраженности в семантике термина компонентов, связанных с семантикой полей. Например, в ключевом слове КОНЦЕПТ СВОЙ–ЧУЖОЙ эксплицитно представлен компонент КОНЦЕПТ, непосредственно входящий в терминополе КОНЦЕПТ И КОНЦЕПТОСФЕРА.

2. В случае отсутствия в семантике термина эксплицитных семантических компонентов, непосредственно входящих в те или иные поля, отнесение термина к полю осуществлялось с опорой на научную традицию, сложившуюся в лингвистике и смежных с ней науках. Например, ЦЕЛОСТНОСТЬ ВОСПРИЯТИЯ ТЕКСТА относится к терминополю РЕЧЕВАЯ ДЕЯТЕЛЬНОСТЬ на основании того, что проблемы восприятия речи обычно рассматриваются в рамках теории речевой деятельности.

3. В случае неоднозначности трактовок КС привлекается контекст его употребления: НКС, название статьи и аннотация.

4. Классификация осуществляется несколькими экспертами; вырабатывается согласованная позиция по спорным вопросам.

5. Одно понятие, выраженное посредством КС, может быть отнесено к одному или к нескольким полям одновременно (см. пример выше).

6. В качестве основного критерия ограничения содержания поля была взята частотность употребления в выборке текстов терминов, входящих в данное поле. В том случае, если частотность терминов, входящих в поле, достигала определенного уровня (определяемого статистически), данное терминополе обретало самостоятельность.

7. При формировании полей авторы стремились: а) по возможности сохранять определенность предметов исследований (т.е. не редуцировать поля до общих обозначений лингвистических направлений и парадигм), б) избегать излишней дробности предметов исследований (для сохранения целостного образа информационного поля).

8. Реализация пп. 6 и 7 приводит к тому, что в представленной ниже графосемантической модели ее отдельные терминополя находятся на разных уровнях иерархии ПрО. Например, поле ПСИХОЛОГИЧЕСКИЕ КАТЕГОРИИ связано с самыми разнообразными сторонами психической жизни человека, изучаемыми психологической наукой. В то же время поля ЛЕКСИКА И ФРАЗЕОЛОГИЯ, ГРАММАТИКА и др. характеризуют либо уровни языковой системы/разделы лингвистики, либо языковые единицы/лингвистические объекты (ТЕКСТ И ДИСКУРС).

Однако, если с позиции формального подхода “неравное” положение терминополей можно рассматривать как недостаток классификационной модели, то, на наш взгляд, оно скорее является достоинством, так как создаваемая модель стремится передать реальное положение вещей, эксплицируя “веса” научных проблем, актуальных в настоящее время в когнитивной лингвистике.

СБОР И ПЕРВИЧНЫЙ АНАЛИЗ МАТЕРИАЛА

Исследование осуществлялось с опорой на *метод графосемантического моделирования*, реализованный в Информационной системе (далее – ИС) “Семограф” (http://semograph.com). Для построения графов использовалось программное средство Gephi (http://gephi.org). В качестве основного инструмента исследования графосемантической модели использовался R – язык программирования высокого уровня, предназначенный для статистической обработки данных (http://www.r-project.org).

Работа в ИС “Семограф” включает в себя создание Проекта (рабочего пространства, в котором осуществляется исследовательский цикл, реализованный в ИС “Семограф”). Фрейм Проекта состоит из следующих элементов:

1. Контекст – в нашем случае описание научной статьи. Контекст включает в себя аннотацию, название статьи и НКС.

2. Корпус контекстов (описание всех публикаций журнала “Вопросы когнитивной лингвистики”).

3. Набор значений, описывающих контекст (научную статью), в том числе:

3.1. Метаданные (тип текстовых данных, позволяющих ввести дополнительную информацию о статье: ФИО автора/авторов статьи; название статьи; научный статус автора/авторов статьи (аспирант, без ученой степени, кандидат наук, доктор наук); год публикации (2008, 2009, 2010, 2011, 2012; 2013); номер журнала (№1, №2, №3, №4); город, в котором проживает автор (спектр наименований).

Особенностью использования метаданных как типа переменных в ИС “Семограф” состоит в возможности множественной фильтрации по значениям метаданных и составления из всего собранного научного контента разнообразных выборок. Например, мы можем составить выборки контекстов научных статей, написанных только аспирантами или только докторами наук из Волгограда, опубликовавшими свои работы в 2009 г., в НКС которых используется термин “концепт” и/или “фрейм” и т.п. ИС “Семограф” предоставляет возможность исследовать семантическое пространство как всего корпуса контекстов, так и отдельной выборки.

3.2. Семантические компоненты (НКС к каждой статье). Например, статью “Фольклорный дискурс: когнитивно-дискурсивное исследование” [15, с. 50] характеризует следующий НКС: ФОЛЬКЛОРНЫЙ ДИСКУРС, КОГНИТИВНО-ДИСКУРСИВНОЕ ИССЛЕДОВАНИЕ, КОММУНИКАТИВНЫЕ СТРАТЕГИИ, ЦЕННОСТИ, ДИСКУРСООБРАЗУЮЩИЕ КОНЦЕПТЫ.

4. Семантические поля – в нашем случае терминополя, объединяющие отдельные научные понятия, выраженные с помощью КС. Терминополе репрезентирует входящие в него термины, что позволяет решить проблему вариативности КС (в том числе и терминологических повторов разного рода, например, синонимических). Для выполнения данного этапа требуется привлечение экспертов моделируемой научной отрасли. В задачу экспертов входит группирование доступных терминов (КС к статьям) в терминополя, формулирование названий терминополей.

Отметим, что связи между полями и компонентами относятся к типу “многие ко многим”, т.е. возможно вхождение одного термина в несколько терминополей и нескольких терминов в одно терминополе. Классификация терминов должна выполняться несколькими экспертами, т.к. существует вероятность субъективного отнесения терминов к терминополям.

5. Графосемантическая модель, объединяющая вышеописанные элементы фрейма Проекта, может дополняться новыми контекстами и допускает изменение семантических полей и их связей с компонентами. На основе графосемантической модели строятся семантическая карта (С-карта) и семантический граф (С-граф).

5.1. С-карта отражает совместное присутствие двух терминополей в одном и том же контексте с учетом подобной встречаемости во всех контекстах выборки. Полагается, что если два КС даются в описании одной и той же статьи, то они становятся связанными между собой через отнесение их к одному контексту. Соответствующим образом мы делаем вывод о связи между терминополями, в которые входят указанные компоненты. С-карта автоматически генерируется на основе подсчета количества связей между полями в пределах всей выборки.

5.2. С-граф представляет собой визуальное представление С-карты.

С-ГРАФЫ ТЕРМИНОВ И ТЕРМИНОПОЛЕЙ ПрО ЖУРНАЛА “ВОПРОСЫ КОГНИТИВНОЙ ЛИНГВИСТИКИ”

Генерация С-карты и С-графа может осуществляться на “сыром” материале: узлами структуры будут являться термины, а ребрами – показатели частоты их совместного использования в НКС. Недостатком такой модели является ее размер: в нашем случае структурная модель состояла бы из 1650 узлов (общее количество КС в корпусе). Поэтому возможности использования С-карты и С-графа терминов ограничены и представляют интерес только в аспекте моделирования наиболее значимого фрагмента структуры ПрО. На рис. 1 представлен С-граф наиболее частотных терминов ПрО КОГНИТИВНАЯ ЛИНГВИСТИКА (порог значимости определялся как превышающий определенную частоту появления термина (в данном случае 0,01) в корпусе).

Представленная модель отражает “взгляд” на ПрО со стороны исследователей ее наиболее разработанного сектора. Однако данная модель передает лишь 8,1 % всех связей между терминами в ПрО. Поэтому для релевантного представления ПрО используется метод полевого анализа, результатом применения которого является структурированная в виде семантической карты (С-карты) и семантического графа (С-графа) *система терминополей*. Кроме того, выявляются зависимости между временными (дата публикации), топологическими (географическими), социальными факторами и моделируемым информационным пространством ПрО. Соотнесенность семантики ПрО с территориальной, социальной, временной “картографией” позволяет понять логику развития ПрО, т.е. проследить пути, способы и механизмы реализации предметной интенции через социальные, географические и временные каналы обмена информацией (и их возможные комбинации).

Полевый анализ ключевых терминов к 446 научным статьям, опубликованным в журнале “Вопросы когнитивной лингвистики” за анализируемый шестилетний период позволил выделить 30 терминополей (табл. 1).

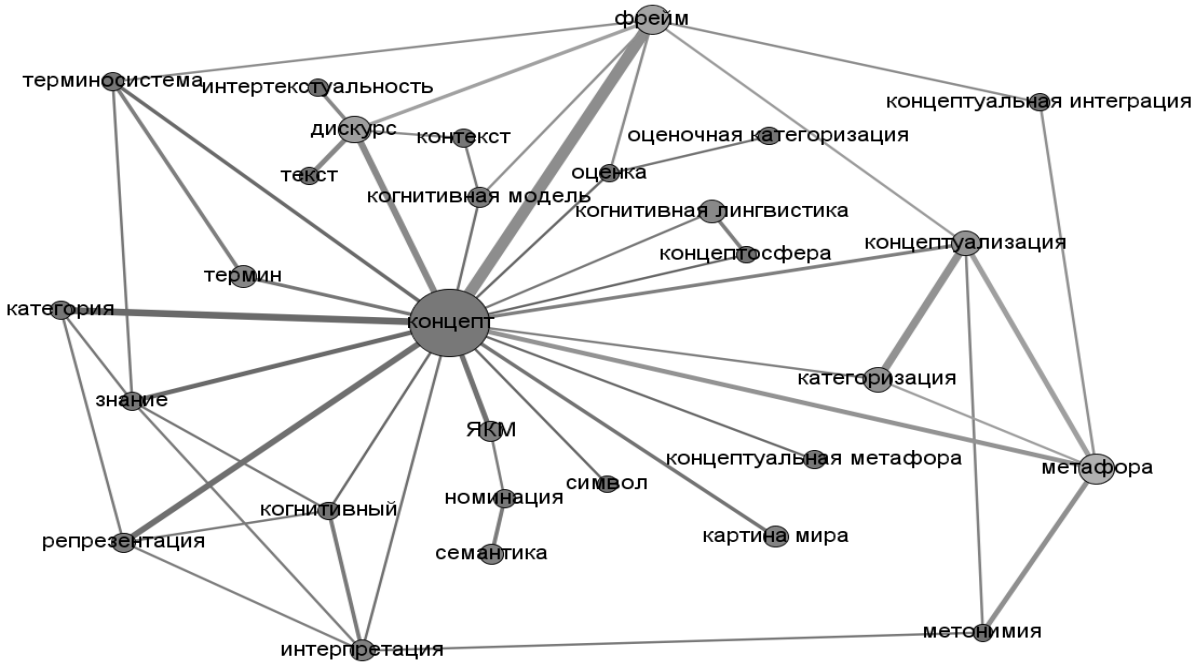

Рис. 1. Графосемантическая модель совместной встречаемости наиболее частотных терминов ПрО КОГНИТИВНАЯ ЛИНГВИСТИКА

Примечание. Размер узла отражает встречаемость КС в проанализированном материале. Толщина линии маркирует силу связи (частоту совместного использования в НКС) между терминами.

Таблица 1

Терминополя и частота их встречаемости в корпусе

Терминополе	Частота	Примеры терминов, входящих в терминополе
Структурная организация	0,633	организация, структура, фрейм, сценарий и мн. др.
Концепт и концептосфера	0,402	концепт, мифоконцепт, концепт дурак и мн. др.
Когниция	0,324	когнитивный, лингвокогнитивный уровень и мн. др.
Сфера культуры и лингвокультуры	0,316	лингвокультурема, мифоконцепт, идеология и мн. др.
Текст и дискурс	0,277	дискурс, интертекстуальность, дискурс-анализ и мн. др.
Психологические категории	0,250	память, экспрессия, ассоциации и мн. др.
Лексика и фразеология	0,245	лексика, идиомы, фразеологизм, термин и мн. др.
Коммуникация	0,215	манипуляция, коммуникативная стратегия и мн. др.
Понятийно-категориальная сфера	0,210	категория, модусные категории, категоризация и мн. др.
Семиотическое пространство	0,205	знак, семиосфера, биосемиотика и мн. др.
Язык	0,197	язык, языковое сознание и мн. др.
Грамматика	0,189	числительное, синтаксис, грамматикализация и мн. др.
Предметная область	0,189	экономика, корпоративная культура, политика и мн. др.
Семантика	0,162	семантические сети, семантическое поле и мн. др.
Тропы и фигуры	0,157	метафора, метонимия, антитеза, символ и мн. др.
Модель и моделирование	0,141	модель, когнитивное моделирование и мн. др.
Направления, школы	0,141	когнитивная лингвистика, теория дискурса и мн. др.
Методы исследования	0,136	шкалирование, сравнительный анализ и мн. др.
Процессуальность	0,130	динамичность, когнитивные процессы и мн. др.
Знание и информация	0,128	знание, информация, форматы знаний и мн. др.
Понимание и смысл	0,125	смысл, когнитивная модель переосмысления и мн. др.
Картина, образ, модель мира	0,117	картина мира, образ мира, модель мира и мн. др.
Речевая деятельность	0,109	речь, речекоммуникативный акт и мн. др.
Значение	0,101	перенос значения, значение, многозначность и мн. др.
Социосфера	0,088	гендер, социальная маркированность и мн.др.
Пространство и время	0,088	время, пространство, замкнутое пространство и мн.др.
Логические категории и операции	0,074	отрицание, противопоставление, аргумент и мн.др.
Сфера аксиологии	0,069	оценка, прагматический оценочный абсурд и мн.др.
Функциональность	0,059	функция, интерпретирующая функция и мн.др.
Языковая личность	0,040	языковая личность, авторская модальность и мн.др.

Примечание: Показатели частоты рассчитываются как отношение веса терминополя к количеству публикаций.

На рис. 2 представлено временное распределение частотности терминополей за период 2008–2013 гг. График дает представление об общем интересе всех авторов журнала “Вопросы когнитивной лингвистики” к тем или иным частным научным направлениям, отражаемым в содержании каждого терминополя.

На рис. 2 видно, что СТРУКТУРНАЯ ОРГАНИЗАЦИЯ – наиболее значимое для ПрО терминополе, что свидетельствует о важности включения в модель всех терминов ПрО, а не только ее наиболее разработанного сектора, в котором доминирующее положение отводилось КОНЦЕПТУ (см. рис. 1).

В то же время информационное пространство каждого конкретного исследования создается на основе взаимодействия нескольких терминополей, что хорошо видно на представленном ниже рис. 3. В качестве примера проанализируем каждое КС из НКС процитированной выше статьи [15]:

ФОЛЬКЛОРНЫЙ ДИСКУРС относится к терминополям КУЛЬТУРА И ЛИНГВОКУЛЬТУРА; ТЕКСТ И ДИСКУРС.

КОГНИТИВНО-ДИСКУРСИВНОЕ ИССЛЕДОВАНИЕ относится к терминополям ТЕКСТ И ДИСКУРС; КОГНИЦИЯ; НАПРАВЛЕНИЯ, ШКОЛЫ.

КОММУНИКАТИВНЫЕ СТРАТЕГИИ относится к терминополям СТРУКТУРНАЯ ОРГАНИЗАЦИЯ; КОММУНИКАЦИЯ.

ЦЕННОСТИ относится к терминополю СФЕРА АКСИОЛОГИИ.

ДИСКУРСООБРАЗУЮЩИЕ КОНЦЕПТЫ относится к терминополям КОНЦЕПТ И КОНЦЕПТОСФЕРА; ТЕКСТ И ДИСКУРС.

Таким образом, в данном наборе ключевых слов актуализовано восемь разных терминополей.

На рис. 3 видно, что минимальное количество полей, актуальных для каждой научной статьи – 1, а максимальное – 13; мода приходится на 6 полей. Данная модель наглядно показывает значимость исследования структурной связности терминополей, формирующих информационное пространство как отдельной публикации, так и журнала в целом, так как структурные связи отражают существующие в реальности комбинации (композиции) научных проблем и направлений.

Моделирование структуры терминополей осуществляется с помощью метода графосемантического моделирования. С-граф терминополей всего информационного пространства журнала “Вопросы когнитивной лингвистики” за шестилетний период 2008–2013 гг. представлен на рис. 4.

Ядро графа – терминополе СТРУКТУРНАЯ ОРГАНИЗАЦИЯ имеет связи почти со всеми терминополями, репрезентируясь в НКС в соответствии с используемыми типами структур в том или ином терминополе. Например, соединение СТРУКТУРНОЙ ОРГАНИЗАЦИИ с КОНЦЕПТОМ И КОНЦЕПТОСФЕРОЙ частотно репрезентируется в терминах ФРЕЙМОВАЯ СТРУКТУРА КОНЦЕПТА, КОНЦЕПТУАЛЬНАЯ МЕЖФРЕЙМОВАЯ СЕТЬ и т.п. В то же время связь СТРУКТУРНОЙ ОРГАНИЗАЦИИ с ГРАММАТИКОЙ актуализирует более традиционные типы структурных наименований: СТРУКТУРА НАРЕЧИЯ, АНАЛИТИЧЕСКИЙ ГЛАГОЛ, СТРУКТУРНАЯ СХЕМА ПРЕДЛОЖЕНИЯ и т.п.

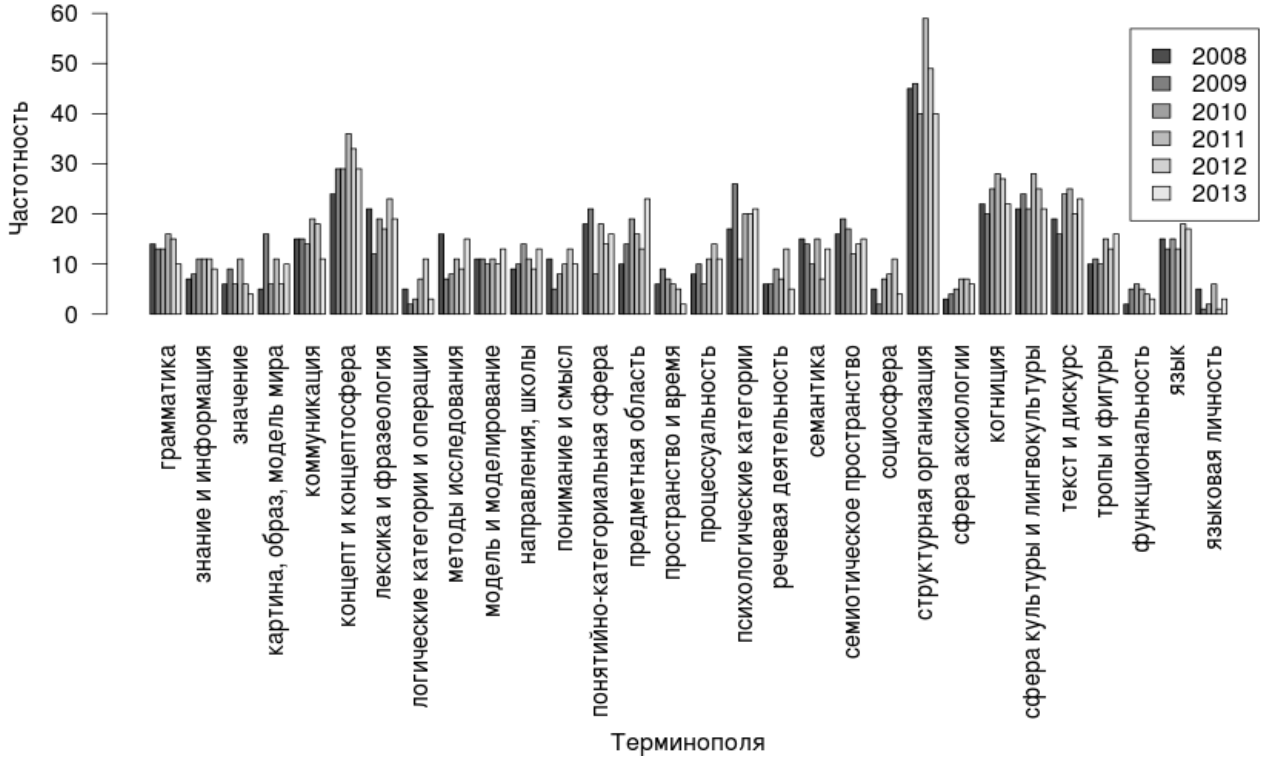

Рис. 2. Распределение частотности терминополей за период 2008–2013 гг.

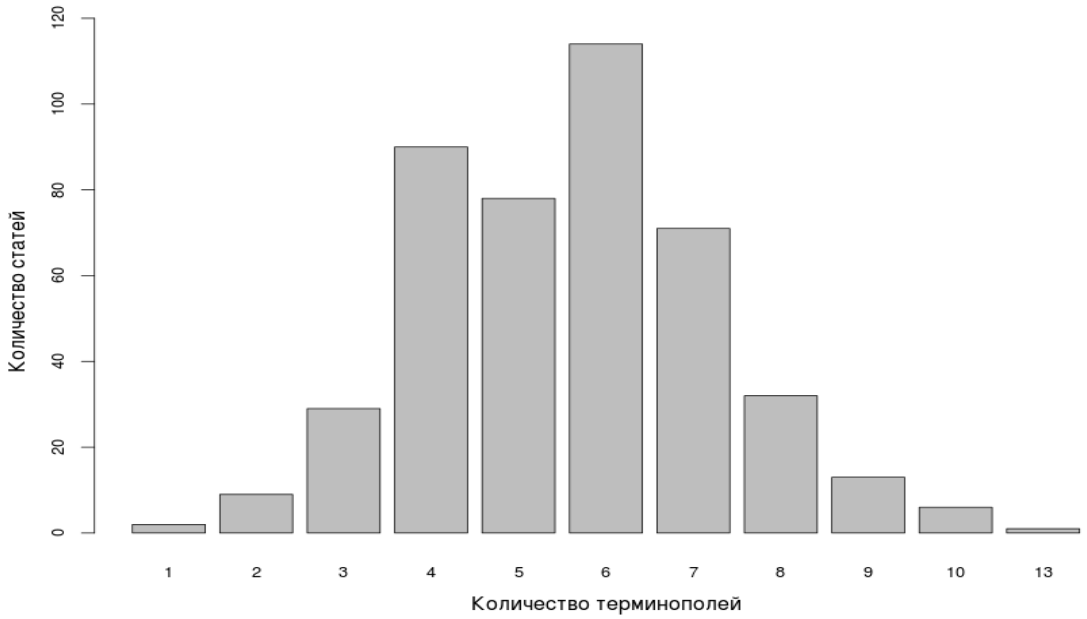

Рис. 3. Распределение количества терминополей по научным статьям за период 2008–2013 гг.

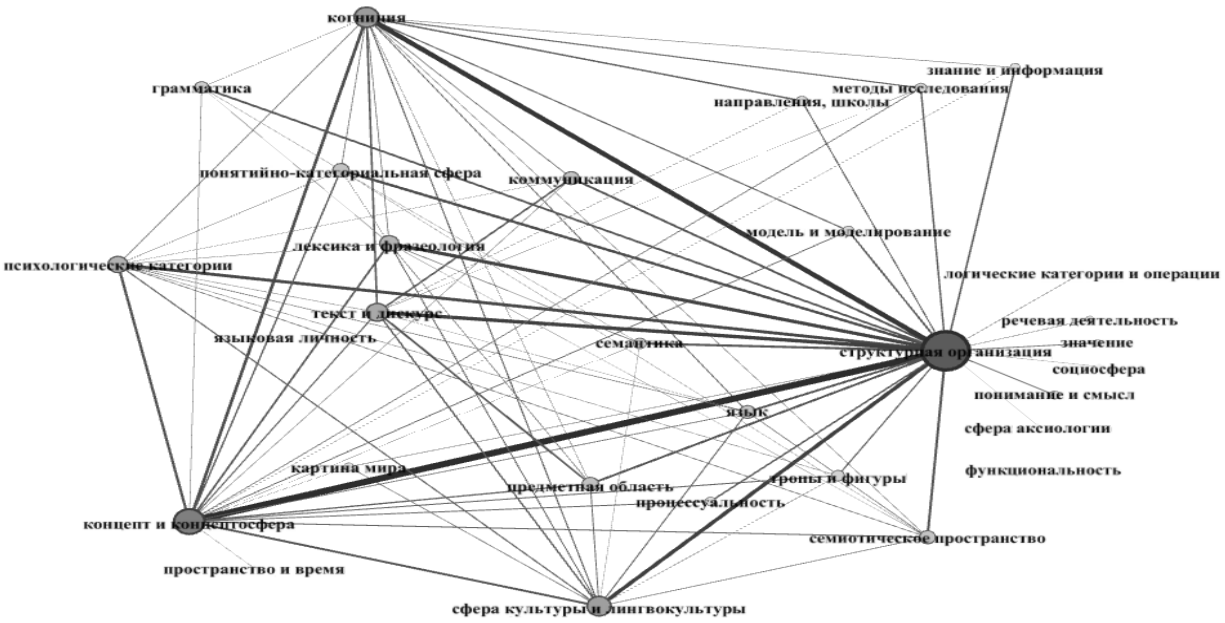

Рис. 4. С-граф терминополей предметной области КОГНИТИВНАЯ ЛИНГВИСТИКА

Примечание. В С-графе вершины соответствуют терминополям, а ребра – связям между ними. Размер вершины пропорционален частотности поля, а толщина ребра – силе связи между полями (частоте совместного присутствия двух терминополей в разных НКС). Порог значимости для ребер превышает частоту 0,05.

На рис. 4 видно, что информационное пространство ПрО в журнале формируется за счет включения в лингвокогнитивную проблематику таких областей, как ТЕОРИЯ ДИСКУРСА и ЛИНГВОКУЛЬТУРОЛОГИЯ, при этом исследовательский “акцент” делается на изучении ЛЕКСИКИ и ФРАЗЕОЛОГИИ (ГРАММАТИКА отходит на второй план). Отметим также значимость терминополя ПСИХОЛОГИЧЕСКИЕ КАТЕГОРИИ, актуального для когнитивистики в целом, однако, отсутствие связей данного терминополя с важными для него терминополями позволяет говорить о недостаточной разработанности психологической проблематики (в частности, отсутствуют связи с терминополями ГРАММАТИКА, КАРТИНА МИРА, ЯЗЫКОВАЯ ЛИЧНОСТЬ и др.).

ЧАСТНОНАУЧНАЯ ПРЕДМЕТНАЯ ОБЛАСТЬ КАК КОМПОЗИЦИЯ ТЕРМИНОПОЛЕЙ

Частнонаучная предметная область – представленная в виде математической модели композиция терминополей, репрезентирующая отдельный научный сегмент в общей ПрО [16, с. 21]. В качестве метода, направленного на выделение частнонаучных ПрО из общей ПрО научного журнала, был использован *кластерный анализ*. Кластерный анализ производится над множеством контекстов, параметрами которых являются бинарные векторы, содержащие наборы полей соответствующих контекстов. Набором полей контекста называется множество терминополей, связанных с данным контекстом посредством произвольного числа терминов (в нашем случае КС к публикациям). Поскольку каждому контексту соответствует частнонаучная ПрО, результат можно использовать в качестве оценки подобия публикаций и соответствующих им частнонаучных ПрО.

Выделение частнонаучных ПрО осуществляется с помощью метода нечеткой кластеризации C-means [17]. В отличие от других распространенных методов кластерного анализа, таких как K-means или самоорганизующиеся карты Кохонена, этот метод допускает принадлежность одного элемента двум и более кластерам, что в данной задаче позволяет отнести одну публикацию к нескольким направлениям в случае высокой неопределенности в ее описании. Кроме того, алгоритм метода C-means допускает возможность влияния на результат с помощью специального числового параметра m, $m \in R$, $m \geq 1$. В качестве исходного значения обычно выбирается m=2, затем, в случае неудовлетворительного результата, выполняется его подстройка. Результаты данного исследования получены с базовым значением параметра m=2.

В результате кластеризации определяются: 1) частнонаучные ПрО, соотносимые с выделенными кластерами, 2) публикации, в той или иной мере соответствующие частнонаучным ПрО.

На основе полученных данных публикации были отсортированы по номеру кластера и расстоянию от его центра в порядке возрастания значений. В табл. 2 представлена кластеризация контекстов в виде 9-кластерной модели, которая была отобрана из спектра других моделей (состоящих из 6, 7, 8, 9 и 10 кластеров) на основании наибольшего соответствия ПрО КОГНИТИВНАЯ ЛИНГВИСТИКА.

Таблица 2

Частнонаучные предметные области (9-кластерная модель)

№ кластера	Композиции терминополей	Условное обозначение	Количество статей	Репрезентативные статьи
1	1. Коммуникация 2. Структура 3. Текст и дискурс	ком-с-тд	66	[18]
2	1. Грамматика 2. Структура	гр-с	69	[19]
3	1. Концепт 2. Лексика и фразеология 3. Структура	к-лф-с	68	[20]
4	1. Направления 2. Структура 3. Когниция 4. Текст и дискурс	нш-с-кгн-тд	41	[21]
5	1. Знание и информация 2. Психол. категории 3. Структура	зн-псх-с	38	[22]
6	1. Концепт 2. Психол. категории 3. Структура	к-псх-с	45	[23]
7	1. Предметная область 2. Культура и лингвокультура 3. Текст и дискурс	по-клт-тд	43	[24]
8	1. Концепт 2. Лексика 3. Предметная область 4. Семиотика 5. Структура	к-лф-по-сп-с	26	[25]
9	1. Структура 2. Культура и лингвокультура 3. Язык	с-клт-яз	50	[26]

Выявленные на основе кластеризации публикаций частнонаучные ПрО представляют собой основные направления исследований в 2008–2013 гг., а представленная на рис. 5 гистограмма распределения выделенных кластеров (частнонаучных ПрО) по временным срезам отражает временную динамику научных интересов авторов журнала. Представленная диаграмма дает возможность оценить, насколько динамичны, изменчивы и в то же время в отдельных аспектах взаимосвязаны области научных интересов журнала.

Так, в частности, увеличение в последние годы количества публикаций кластера 3 “к-лф-с” (реконструкция концептов на основе лексического анализа) свидетельствует о выборе журналом в качестве приоритетных проблем концептологии. Альтернативное направление (кластер 6 “к-псх-с”), имеющее в 2008 г. близкую стартовую позицию, в дальнейшем вытесняется на периферию (особенно это заметно по количественным данным 2013 г.).

Полагаем, что вытеснению данного направления способствовали набирающие популярность исследования в рамках кластера 8 (к-лф-по-сп-с), связанные с исследованием терминосистем и концептосфер специального знания. Аналогичным образом связаны между собой исследования в области теории дискурса (кластеры 1, 4 и 7) и лингвокультурологии (кластеры 7 и 9).

На рис. 6 представлена карта научных интересов авторов журнала (сгруппированных по географической принадлежности), составленная в соответствии с выделенными частнонаучными ПрО.

Рис. 6 дает наглядное представление, с одной стороны, о научных интересах авторов, а с другой – об обоснованности самой группировки по городам, показывающей “научную географическую специализацию”. В том случае, когда речь идет о ведущих научных центрах (Москва, Санкт-Петербург) и достаточно большом объеме публикаций, сферы научных интересов распределяются почти равномерно. Небольшие научные центры даже при внушительном объеме публикаций имеют тенденцию к специализации. Особенно это заметно с публикациями исследователей из Махачкалы (почти все находятся в кластере 2) и статьями представителей Тверской лингвистической школы.

В то же время изучение научной “географической специализации” нужно осуществлять с временной “привязкой” публикаций, поскольку это позволит: 1) сопоставить синхронные изменения в информационном пространстве ПрО (на рис. 6 сравниваются работы, созданные в разные периоды времени); 2) выявить наличие/отсутствие изменений в исследовательских интересах к тем или иным научным проблемам; 3) установить зависимости в появлении интересов к тем или иным научным проблемам.

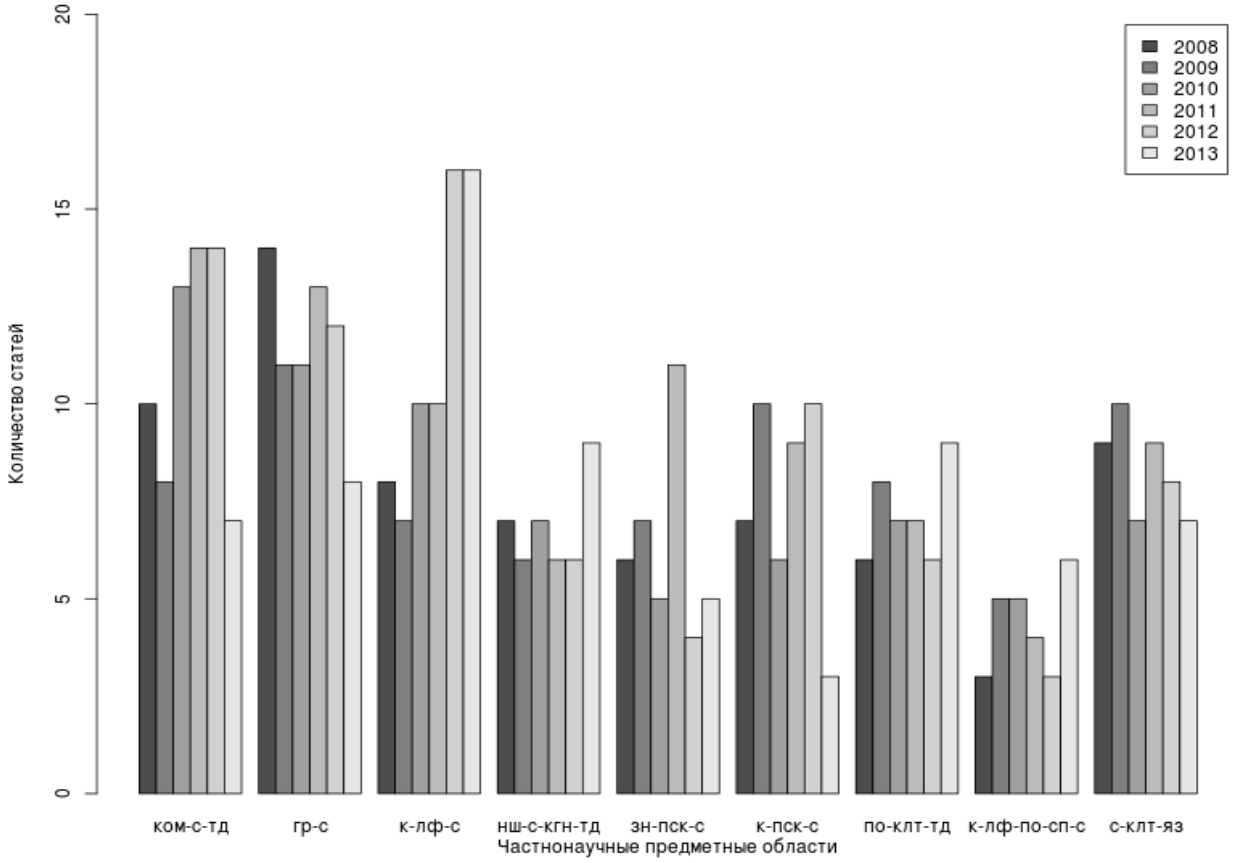

Рис. 5. Распределение выделенных кластеров (частнонаучных предметных областей) по временным срезам 2008–2013 гг.

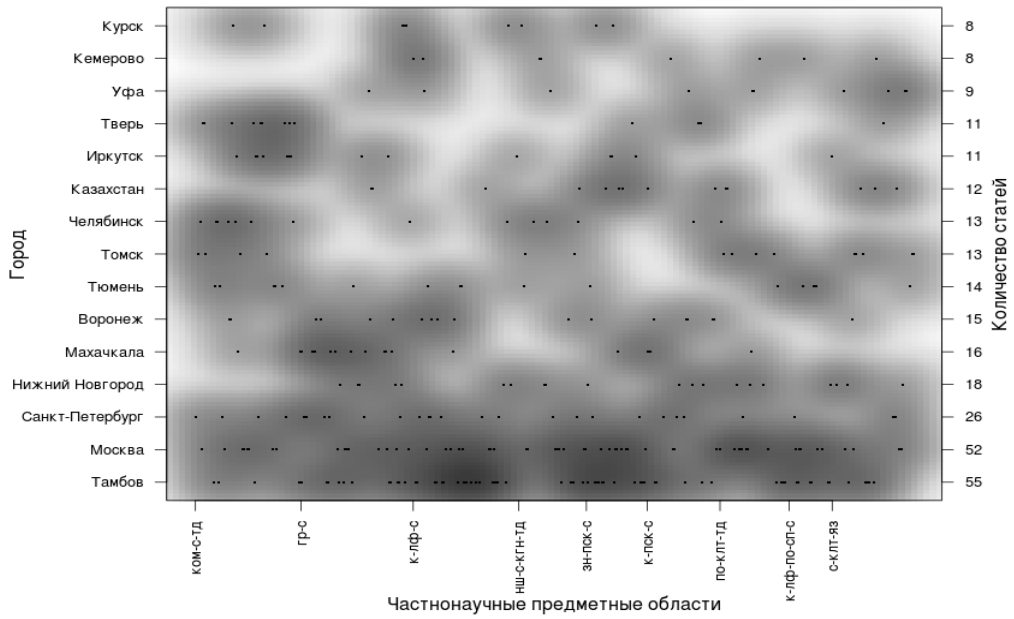

Рис. 6. Карта научных интересов авторов журнала "Вопросы когнитивной лингвистики" за период 2008–2013 гг. (по частнонаучным предметным областям)

Примечание. Точками обозначены отдельные публикации. Центры кластеров располагаются на линиях, проведенных вертикально из намеченных кластеров (над их названиями). Чем ближе расположена точка к такой вертикальной линии, тем более репрезентативна для данного кластера публикация.

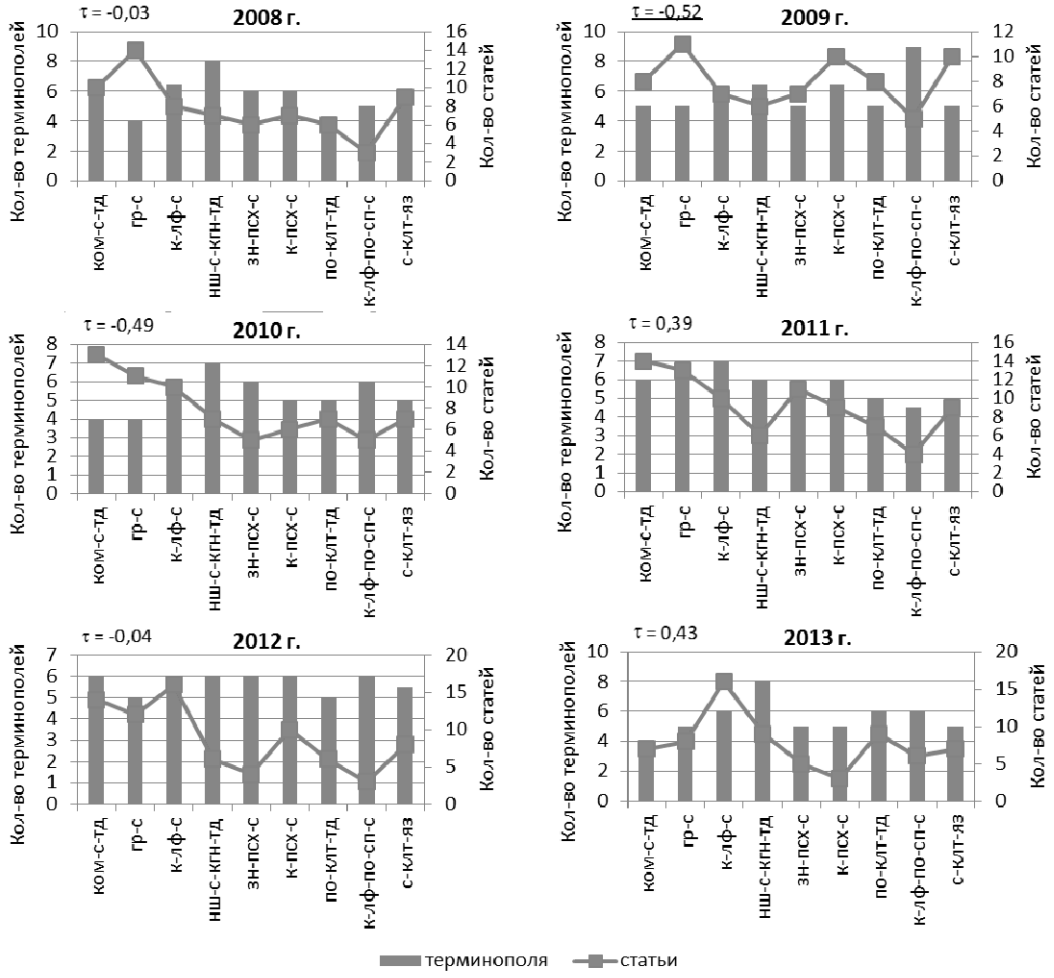

Рис. 7. Временная динамика распределения по кластерам среднего количества терминополей в НКС к публикациям и количества публикаций (группировка по годам).

Примечание. Подчеркнуты показатели значимых корреляций при $p<0,05$.

Имеющиеся данные позволяют рассматривать проблему в следующем аспекте: 1) по каждой частнонаучной ПрО есть данные, свидетельствующие об интересе к ней в любом временном срезе (см. рис. 5); 2) в то же время имеются данные о среднем количестве терминополей, актуализованных в НКС к статьям каждой частнонаучной ПрО в тех же временных срезах. Можно предположить, что разработка частнонаучной ПрО будет приводить к ее усложнению, что на формальном уровне может выражаться в увеличении среднего количества терминополей в НКС к статье. А временная динамика среднего количества терминополей в НКС к статье в частнонаучной ПрО должна соотноситься с количеством публикаций по данной тематике (т.е. с интересом к изучаемому проблемному полю).

На рис. 7 и 8 соотнесены два ряда данных – среднее количество терминополей в НКС к статье и количество публикаций. Рис. 7 представляет данные, сгруппированные по годам, рис. 8 – по выделенным кластерам. На каждом рисунке отображаются показатели коэффициента корреляции Кендалла (показатели значимых корреляций при $p<0,05$ подчеркнуты).

На рис. 8 видно, что только три частнонаучные ПрО имеют корреляцию уровня сложности ПрО с ее популярностью. Однако со временем корреляции между сложностью и популярностью увеличиваются (см. рис. 7): от отсутствия корреляции (2008 г.) до отрицательной корреляции (2009 и 2010 гг.) и изменения знака на положительную корреляцию (2011 и 2013 гг.). Обнаруженные закономерности между уровнем сложности/разработанности ПрО и ее популярностью среди исследователей нуждаются в более масштабной проверке и в случае подтверждения могут использоваться для оценки, прогноза и планирования редакционно-издательской деятельности научных журналов или для дополнительного обоснования требуемого уровня поддержки научными фондами тех или иных исследовательских направлений (о существующих методах см., в частности, [27]).

Для прогнозирования значений количества терминополей и публикаций в 2014 и 2015 гг. воспользуемся методом Multiple Regression (Множественная регрессия), реализованном в статистическом пакете Statistica 8.0 (см., например, [28, с. 180–190]). Прогнозируемые показатели рассчитываются для каждого выделенного кластера; результаты представлены на рис. 8. На рис. видно, что прогнозируется наибольший рост количества публикаций кластера 3 “к-лф-с” и уменьшение количества публикаций конкурирующих с ним кластеров 6 “к-псх-с” и 8 “к-лф-по-сп-с”. Между тем, кластер 3 – наиболее простая частнонаучная ПрО (достаточно сказать, что она значительно чаще остальных используется в квалификационных работах). Кластеры 6 и 8 представляют, на наш взгляд, большую теоретическую и прикладную ценность, поэтому складывающаяся ситуация требует некоторого “выравнивания” частнонаучных ПрО, придания импульса наиболее значимым из них.

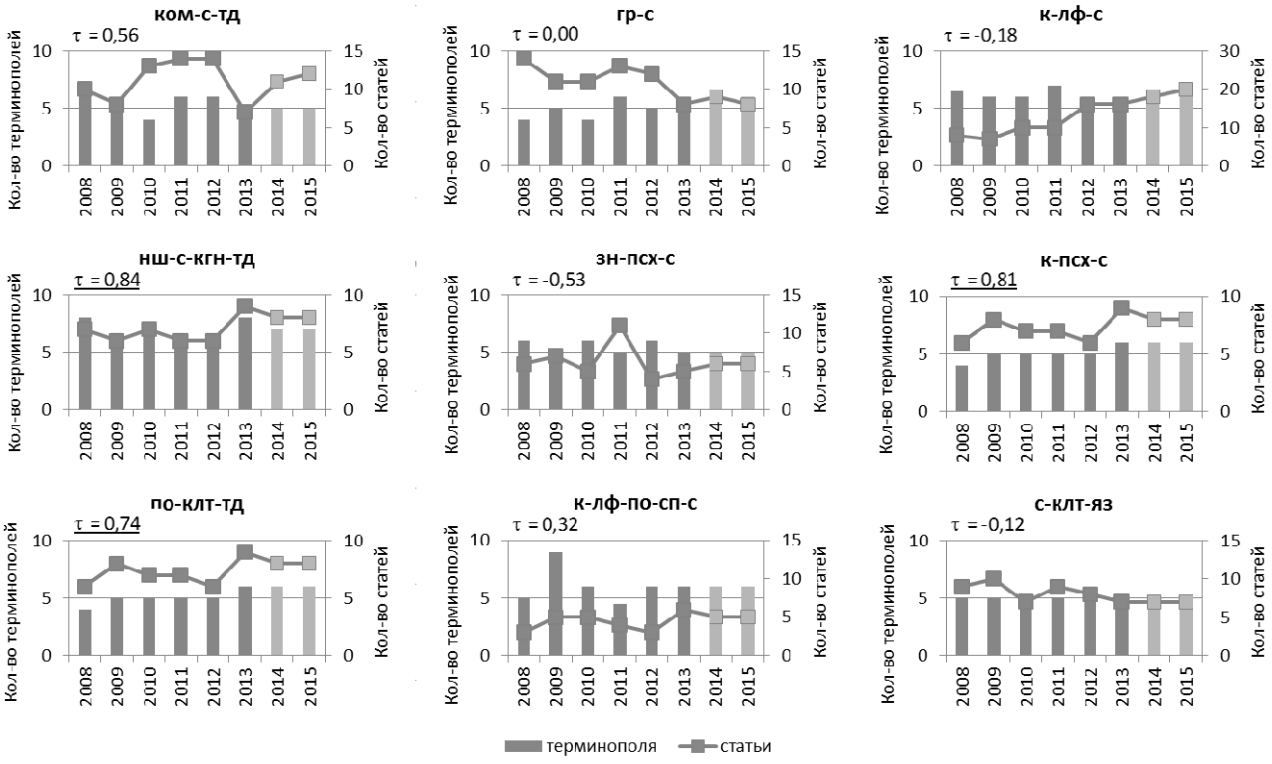

Рис. 8. Временная динамика распределения (с прогнозом) по кластерам среднего количества терминополей в НКС к публикациям и количества публикаций (группировка по кластерам)

Примечание. Подчеркнуты показатели значимых корреляций при $p<0,05$ (прогнозируемые значения не учитывались).

ВЫВОДЫ

Изучение репрезентации ПрО в тематическом научном журнале представляет интерес в самых разных аспектах. Анализ ПрО с позиции "живой науки" позволяет выявить частнонаучные ПрО, определить их "веса" и динамику изменений, а также представить прогнозные модели развития. Представленные результаты демонстрируют возможности используемого метода исследования и позволяют говорить о концептосфере научного журнала как о постепенно оформляющейся системе терминополей; при этом сам процесс оформления концептосферы (ПрО) в отдельных ее секторах имеет разную динамику. Отчасти данная динамика обусловлена социальными факторами, в частности, научным статусом участников (можно отметить более высокую корреляцию между динамикой сложности частнонаучной ПрО и количеством публикаций в случае большего участия в ее разработке докторов наук). Вероятно, влияние на динамику отдельных частнонаучных ПрО оказывают также широта географического распространения, традиции научных школ, возможности междисциплинарного синтеза.

Полагаем, что существующие методы и теоретический базис терминоведения, рассмотренные в контексте современных информационных технологий и методов математического моделирования, позволяют решать не только классификационные задачи, но и задачи оценки, прогнозирования и управления исследовательской деятельностью, в том числе и в рамках создания систем поддержки принятия решений в данной сфере.

СПИСОК ЛИТЕРАТУРЫ

1. Nederhof A.J., van Leeuwen T.N., van Raan A.F.J. Highly cited non-journal publications in political science, economics and psychology: a first exploration // Scientometrics. – 2010. – Vol. 83, Iss. 2. – P. 363–374.
2. Григорьев В.К. Научные журналы – механизм объективной оценки грантополучателей в технологии закрепления вузовской молодежи в науке // Дистанционное и виртуальное обучение. – 2013. – № 4. – С. 110–119.
3. Álvarez-de-Toledo-Saavedra L. Bibliographic control and dissemination of the University of Oviedo scientific output // El profesional de la información. – 2012. – Vol. 21, Iss. 6. – P. 639–642.
4. Стародубов В.И., Куракова Н.Г., Цветкова Л.А., Арефьев П.Г., Кураков Ф.А. Проблемы оценки мирового уровня конкурентоспособности российской науки на примере национальной клинической медицины // Научно-техническая информации. Сер.1. – 2012. – № 8. – С. 1–15; Starodubov V.I., Kurakova N.G., Tsvetkova L.A., Aref'ev P.G., Kurakov A.F. The Problems Associated with the Evaluation of World-Class Competitiveness of Russia's Science, as Illustrated by Clinical Medicine // Scientific and Technical Information Processing. – 2012. – Vol. 39, № 3. – P. 139–152.
5. Москалева О.В. Можно ли оценивать труд ученых по библиометрическим показателям? // Управление большими системами. Специальный выпуск 44: Наукометрия и экспертиза в управлении наукой. – 2013. – С. 308–331.
6. Базы данных ИНИОН. – URL: http://www.inion.ru/ index6.php (дата обращения: 18.04.2014).
7. Мдивани Р.Р. Тезаурусы ИНИОН РАН по социальным и гуманитарным наукам // Научно-техническая информация. Сер. 1. – 2013. – № 7. – С. 23–27.
8. Arano S. Thesauruses and ontologies // Hipertext.net. № 3. – URL: http://eprints.rclis.org/8972/2/12.pdf (дата обращения: 18.04.2014).
9. Wielinga B.J., Schreiber A.Th., Wielemaker J., Sandberg J.A.C. From thesaurus to ontology // K-CAP '01: Proceedings of the 1st international conference on Knowledge capture. – ACM New York, 2001. – P. 194–201.
10. Гладун А.Я., Рогушина Ю.В. Использование онтологических знаний и тезаурусов для объективного профилирования специалистов // Искусственный интеллект. – 2006. – № 3. – С. 379–390.
11. Никитина С.Е. Семантический анализ языка науки. На материале лингвистики. – М.: Наука, 1987. – 146 с.
12. Кузнецов А.М. От компонентного анализа к компонентному синтезу. – М.: Наука, 1986. – 126 с.
13. Морозова Л.А. Терминознание: Основы и методы. – М.: ГНО "Прометей" МПГУ, 2004. – 144 с.
14. Щур Г.С. Теории поля в лингвистике. – М.: Эдиториал УРСС, 2009. – 264 с.
15. Эмер Ю.А. Фольклорный дискурс: когнитивно-дискурсивное исследование // Вопросы когнитивной лингвистики. – 2011. – № 2. – С. 50–60.
16. Белоусов К.И., Баранов Д.А., Зелянская Н.Л. Научный коллектив и его предметные области (к вопросу о методах эффективного планирования научной деятельности) // Научно-техническая информация. Сер. 1. – 2014. – № 4. – С. 13–26.
17. Kaymak U., Setnes M. Extended fuzzy clustering algorithms // Erasmus Research Institute of Management. – 2000. – URL: http://repub.eur.nl/pub/57/erimrs20001123094510.pdf (дата обращения: 18.04.2014).
18. Виноградова С.А. Инструменты речевой манипуляции в политическом медиадискурсе // Вопросы когнитивной лингвистики. – 2010. – № 2. – С. 95–101.
19. Березина О.А. К вопросу о языковой репрезентации категории безличности // Вопросы когнитивной лингвистики. – 2010. – № 4. – С. 39–51.
20. Непрокина Ю.А., Пискунова С.В. Концепт ЛЕТО и его воплощение в языковой культуре тамбовского края // Вопросы когнитивной лингвистики. – 2009. – № 4. – С. 123–128.

21. Белоусов К.И., Зелянская Н.Л. Лингвосемиотическое моделирование обыденной географической картины мира // Вопросы когнитивной лингвистики. – 2013. – № 2. – С. 73–85.
22. Новикова М.Г. Возможность графической интерпретации процесса понимания // Вопросы когнитивной лингвистики. – 2011. – № 3. – С. 132–137.
23. Коваленко Г.Ф. Межтекстовое ассоциативно-смысловое поле как способ представления художественного концепта LOVE в идиостиле И. Стоуна // Вопросы когнитивной лингвистики. – 2012. – № 3. – С. 76–83.
24. Пеньков Б.В. Ядро и периферия образовательного дискурса (на материале британской средней школы) // Вопросы когнитивной лингвистики. – 2010. – № 3. – С. 84–91.
25. Зыкова И.В. Семиотика музыки в построении фразеологического значения (лингвокультурологический подход) // Вопросы когнитивной лингвистики. – 2012. – № 4. – С. 108–117.
26. Тахтарова С.С. Концепт TOLERANZ в немецкой лингвокультуре // Вопросы когнитивной лингвистики. – 2008. – № 1. – С. 64–71.
27. Рудцкая Е.Р., Хрусталев Е.Ю., Цыганов С.А. Методы накопления научного знания для инновационного развития российской экономики (опыт РФФИ) // Проблемы прогнозирования. – 2009. – № 3. – С. 134–139.
28. Вуколов Э.А. Практикум по статистическим методам и исследованию операций с использованием пакетов Statistica и Excel. – М.: ФОРУМ, 2008. – 464 с.

Материал поступил в редакцию 09.07.14.

Сведения об авторах

БЕЛОУСОВ Константин Игоревич – доктор филологических наук, профессор кафедры теоретического и прикладного языкознания Пермского государственного национального исследовательского университета
e-mail: belousovki@gmail.com

БАРАНОВ Дмитрий Александрович – аспирант кафедры математического обеспечения информационных систем Оренбургского государственного университета
e-mail: baranov@semograph.com

ЕРОФЕЕВА Елена Валентиновна – доктор филологических наук, профессор кафедры теоретического и прикладного языкознания Пермского государственного национального исследовательского университета
E-mail:elevaer@gmail.com

ЗЕЛЯНСКАЯ Наталья Львовна – кандидат филологических наук, ведущий научный сотрудник кафедры теоретического и прикладного языкознания Пермского государственного
национального исследовательского университета
E-mail: zelyanskaya@gmail.com

ИЧКИНЕЕВА Дилара Ахметовна – кандидат филологических наук, доцент кафедры иностранных языков естественных факультетов Башкирского государственного университета
E-mail: dilaraichkineeva@gmail.com

УДК 81’322.2

В.А. Яцко

Метод зонально-корреляционного анализа текста

Описывается метод зонально-корреляционного анализа текста, предполагающий сопоставление распределения слов в зонах J_1 двух или более текстов. Сопоставляемые тексты делятся на зоны J_0, J_1, J_2 в соответствии с разработанной ранее интерпретацией закона Брэдфорда. Проводится сравнительный анализ параметров слов, входящих в зоны J_1 текстов, и вычисляется расстояние, указывающее на степень их смысловой близости. Метод может применяться для автоматической классификации и авторской атрибуции текстов, он позволяет получать адекватные результаты при минимальном количестве параметров

Ключевые слова: *автоматическая классификация текстов, зонально-корреляционный анализ, параметры, степень смысловой близости*

Термин *лингвистическая информатика* был нами предложен [1] для обозначения дисциплины, изучающей закономерности распределения текстовой информации, проблемы, принципы, методы и алгоритмы разработки лингвистического программного обеспечения и аппаратных средств.

Одним из наиболее известных и широко применяемых в предметной области лингвистической информатики законов является закон Ципфа, устанавливающий зависимость между рангом слова и частотностью его встречаемости в тексте [2], которая может быть выражена формулой

$$F(w_{ij}) = F(w1_j) / R(w_{ij}) \quad , \qquad (1)$$

где F – частотность слова w_i в тексте j-м; R – его ранг в ранжированном списке, а $w1$ – слово с первым рангом. Анализ Брауновского корпуса показал, что распределение слов по частотностям и рангам в этом корпусе соответствует закону Ципфа [3] Слово с первым рангом повторяется около 7000 раз, частотность слова со вторым рангом составляет примерно 1/2 от частотности первого слова (≅3500), частотность третьего по рангу слова – 1/3 от частотности первого слова и т.д. Соответственно, произведение частотности определенного слова на его ранг будет постоянной величиной, равной частотности первого по рангу слова, поэтому закон Ципфа записывается также в виде:

$$F * r = C \qquad (2)$$

Этот закон имеет предсказательную силу: зная ранг и частотность одного элемента ранжированного списка, нетрудно определить частотности всех остальных элементов. Если, допустим, десятое по рангу слово повторяется в определенном тексте 36 раз, то частотность пятого по рангу слова будет равна 36*10/5=72.

Дж. Ципф также отметил, что почти половина слов в тексте встречается лишь один раз, 1/6 часть слов – два раза, 1/2 – три раза. Такое распределение по частотностям называют вторым законом Ципфа [4; 5, с. 234].

Для автоматической обработки текста может применяться формула Шеннона:

$$H = -\sum_{i=1}^{n} P_i * \log_2 P_i \ , \qquad (3)$$

где H – количество информации на единицу текста, P_i – вероятностная величина, которая находится делением частотности данной единицы текста на сумму частотностей всех единиц текста, и значение которой изменяется в диапазоне от нуля до единицы, т.е.

$$P(w_{ij}) = \frac{F(w_{ij})}{\sum F(w_j)} \ , \qquad (4)$$

где F – частотность, а w_i – единица текста j-го. В качестве единиц текста могут выступать отдельные символы или слова. В [6, с. 59-60] показано, что количество информации, приходящейся на символ русского алфавита и пробел, H≅ 4,72 бита. При этом вероятностные величины конкретных символов, вычисленные на материале больших по объёму текстов, составили: P(пробел)=0.175, P(о)= 0.90, P(е)=0.72 и т.д.

Намного менее известен и практически мало применяется для автоматической обработки текста закон Брэдфорда, устанавливающий зависимость между зонами распределения научных статей по журналам, которую можно представить в виде

$$S(J_0) : S(J_1) : S(J_2) = 1 : q : q^2, \qquad (5)$$

где S – некоторое количественное значение зон J_0, J_1, J_2, а q – коэффициент Брэдфорда, представляющий собой постоянную величину.

В [7] нами была предложена интерпретация закона Брэдфорда в терминах геометрической прогрессии (*Y*-интерпретация):

$$S_n = J_2(q^n - 1)/(q - 1), \qquad (6)$$

где $S_n = S(J_0) + S(J_1) + S(J_2)$; q – коэффициент Брэдфорда (зональный коэффициент), задающий соотношение между зонами; n= 3 – постоянная величина, равная количеству зон.

Y-интерпретация существенно отличается от закона Брэдфорда, поскольку отсутствует отношение включения между источниками информации (журналами) и самой информацией (статьями), что позволяет значительно расширить сферы применения этой интерпретации. Другим отличием является возможность изменения значения n, и, соответственно, количества зон, которое может быть любым, в том числе дробным. Предложенная интерпретация может применяться для определения пороговых уровней при выделении некоторых подмножеств из множества, например, подмножества терминов в тексте или групп регионов в зависимости от результатов голосования.

Цель настоящей статьи – описать метод зонально-корреляционного анализа текста, разработанный нами на основе *Y*-интерпретации закона Брэдфорда и предполагающий сопоставление зон двух или более текстов с целью их классификации, а также авторской атрибуции.

Деление текста на зоны на основе предложенного закона предусматривает следующие этапы.

1. Построение числового ряда и вычисление его суммы (С). Под числовым рядом понимается список единиц текста (как правило, слов), в котором каждой единице приписано числовое значение. Список сортируется по нисходящей.

2. Вычисление количественного значения зоны $S(J_2)$. J_2 – зона с наибольшим рассеянием информации, содержащая большое количество источников с наименьшим количеством информации, приходящейся в среднем на один источник. Количественное значение этой зоны (S) вычисляется по формуле:

$$S(J_2) = C / K, \qquad (7)$$

где $K = (q^n - 1)/(q - 1)$; n = 3 (константа, равная количеству зон); q – коэффициент, определяемый эмпирически, специфичный для каждой конкретной предметной области. Нетрудно убедиться, что при q = 3 $S(J_2)$=С / K=С /13 ($(3^3 - 1)/(3 - 1) = 13$), а при q = 2 $S(J_2)$=С / K =С /7. В [7] приводится таблица возможных коэффициентов K при возможных значениях q.

3. Вычисление количественного значения зон J_0,J_1, которое производится последовательным умножением на коэффициент q, т.е.:

$$S(J_1) = S(J_2)*q;\ S(J_0) = S(J_1)*q\ . \qquad (8)$$

4. Определение количественного состава трёх зон, т.е. сколько и какие именно единицы текста входят в каждую зону. Для этого с помощью табличного процессора изменяется диапазон количественного значения каждой из зон до тех пор, пока не найдётся значение, наиболее близкое к ранее определённому количественному значению каждой зоны. Это значение и будет пороговым уровнем, отделяющим одну зону от другой.

Допустим, существует список единиц текста W1….W50, в котором каждой единице соответствует некоторое число:
W1=68, W2=51, W3=37, W4=28, W5=25, W6=23, W7=20, W8=18, W9=17, W10=15, W11=15, W12=13, W13=12, W14=12, W15=12, W16=12, W17=11, W18=11, W19=10, W20=10, W21=10, W22=10, W23=10, W24=9, W25=9, W26=8, W27=8, W28=7, W29=7, W30=7, W31=7, W32=7, W33=7, W34=6, W35=6, W36=6, W37=6, W38=6, W39=6, W40=6, W41=6, W42=6, W43=6, W44=6, W45=6, W46=6, W47=6, W48=6, W49=5, W50=5.

Вначале находится сумма всех количественных значений, C = 68+51+37…+5 = 626. Далее по формуле (7) находится значение зоны $S(J_2) = C / K = C / = C$ / 13 = 626 / 13 = 48,15385. Это число соответствует сумме количественных значений m элементов из нижней части списка, начиная с W50. Затем по формуле (2) находятся количественные значения для двух других зон: $S(J_1)$= S(J_2)*q = 48,15385 * 3 = 144,4615; $S(J_0)$= $S(J_1)$*q = 144,4615 * 3 = 433,3846. Для определения состава зоны J_2 задаём произвольный диапазон, начиная с последнего члена числового ряда W50. Например, если задать диапазон 15, то сумма 15-ти последних членов числового ряда (W36:W50) будет равна 88, что существенно больше необходимой величины (абстрактного порогового уровня) $S_a(J_2)$=48,15385. Соответственно, уменьшаем диапазон, пока не находим значение (S_r – реальный пороговый уровень), максимально близкое к $S_a(J_2)$. Таким значением будет $S_r(J_2)$=46 при диапазоне W43:W50 (отклонение от $S_a(J_2)$ = 2,15385). Если диапазон увеличить до W42:W50, то получим $S(J_2)$=52, что значительно отклоняется от абстрактного порогового уровня (отклонение = 4,15385). Если диапазон уменьшить до W44-W50, то отклонение составит $S_a(J_2)$= 48,15385 – $S_r(J_2)$ =40 = 8,15385.

Таким образом, в зону J_2 войдут 8 объектов. Аналогично находим количественный состав зоны J_0 (т.е. $S(J_0)$), изменяя диапазон с первого члена числового ряда W1. После его нахождения оставшиеся объекты попадут в зону J_1 (табл. 1).

Таблица 1

Результаты вычислений количественного состава зон $J_0 - J_2$

Зона	S_a	S_r	Диапазон	Количество объектов
J_0	433,3846	430	W1:W21	21
J_1	144,4615	150	W22:W42	21
J_2	48,15385	46	W43:W50	8

В этом примере (см. табл. 1) сначала было найдено значение S(J_2), затем – S(J_0), а затем – S(J_1), т.е. анализ проводился в последовательности

(1) $S(J_2) \rightarrow S(J_0) \rightarrow S(J_1)$.

Однако возможны и другие варианты:

(2)$S(J_2) \rightarrow S(J_1) \rightarrow S(J_0)$;

(3)$S(J_0) \rightarrow S(J_1) \rightarrow S(J_2)$.

Выбор той или иной последовательности зонального анализа определяется его целями: в первую очередь вычисляется значение той зоны, для которой нужно получить наиболее точный результат. В вычислениях, которые будут приведены далее, используется вариант (3), поскольку для настоящего исследования важно получить наиболее точные данные для $S(J_0)$ и $S(J_1)$, а $S(J_2)$ будет определяться по "остаточному" принципу.

Для того чтобы показать применение зонального анализа на примере конкретного текста, был создан файл, состоящий из пяти романов Т. Дразера: «*The Genius*», «*The Financier*», «*Titan*», «*Sister Carrie*» и «*Genie Gerhardt*», которые были взяты с сайта Проекта Гутенберг.[1] Из текстов была удалена информация о проекте, в результате получился текст, включающий 23591 уникальное слово и 1003944 токена; эти данные были получены с помощью программы *AntConc concordancer* [8]. Размер текста (около миллиона токенов) соответствует размеру Брауновского корпуса, распределение слов в котором подчиняется закону Ципфа. Работы Т. Драйзера характеризуются специфическим натуралистическим стилем, что обеспечивает более наглядные результаты при сопоставлении с работами других авторов.

Результаты зонального анализа, проводившегося при q=3, представлены в табл. 2. Ранее нами было показано, что именно это значение зонального коэффициента является адекватным при анализе текстов, в то время как при анализе других типов данных могут использоваться другие значения коэффициента [9].

Результаты анализа могут быть интерпретированы в терминах второго закона Ципфа. В зону J_0 входят слова, встречающиеся с высокой частотностью в текстах, к которым относятся стоп-слова, т.е. служебные слова, такие как артикли, предлоги, местоимения, союзы. Зону J_2 составляют слова, редко используемые автором, а также редко встречающиеся в текстах, например, собственные имена, авторские неологизмы, сокращения, жаргонизмы. Зона J_1 включает слова, представляющие основное содержание текста; распределение этих слов по частотностям характеризует специфику данного текста или группы текстов.

Соответственно, зональный анализ может быть использован для фильтрации стоп-слов, информационного поиска, автоматической классификации и авторской атрибуции текстов. Последняя задача обычно решается на основе сопоставления распределения некоторых параметров в двух или более текстах и определения расстояния между этими текстами с помощью векторного моделирования документов [10]. Тексты, расстояние между которыми является наименьшим, относятся к одной группе, жанру или классу. Зональный анализ позволяет проводить сопоставление распределения параметров в отдельных зонах текстов, а не в текстах в целом. Такое сопоставление мы называем зонально-корреляционным анализом. Объектом сопоставления, в первую очередь, должны быть зоны J_1, поскольку содержащиеся в них слова представляют основное содержание текста.

Таблица 2

Зональный анализ текстов Т.Драйзера

Зона	Абстрактный пороговый уровень	Реальный пороговый уровень	Диапазон	Количество слов	Первые шесть слов	Частотности
J_0	695038.1538	694910	1 - 287	287	the	40904
					to	29792
					and	28970
					of	26706
					he	24383
					a	21711
J_1	231679.3846	231681	288 -3472	3185	moment	386
					kind	385
					state	384
					white	384
					situation	383
					ought	382
J_2	77226.46154	77353	3473 - 23591	20119	fortunes	20
					funny	20
					fuss	20
					gaiety	20
					generosity	20
					greeting	20
Всего	**1003944**		**23591**			

[1] http://www.gutenberg.org/wiki/Main_Page

Применение зонально-корреляционного анализа можно продемонстрировать, сопоставив проанализированные выше тексты Т. Драйзера (файл Dreiser1) с текстами романов Ч. Диккенса «*Oliver Twist David Copperfield*», «*Bleak House*» и «*The Mystery of Edwin Drood*» (файл Dickens), а также с текстами других работ Т. Драйзера: «*An American Tragedy*», «*The Stoic*», «*Free and Other Stories*» и «*Twelve Men*» (файл Dreiser2). Расстояние между работами разных авторов (Dreiser1 и Dickens), должно быть существенно больше, чем между работами одного автора (Dreiser1 и Dreiser2), расстояние между которыми будет намного меньше.

Вычисление этого расстояния в процессе зонально-корреляционного анализа предусматривало разделение текстовых файлов на три зоны по методике зонального анализа, а затем выполнение пересечения зоны J_1 файла *Dreiser1* с соответствующими зонами файлов *Dickens* и *Dreiser2*. Так образовались зоны пересечения *A* и *B*. В зону *A* вошли слова, которые встречаются как в зоне J_1 файла *Dreiser1*, так и в зоне J_1 файла *Dickens*. В зону *B* вошли слова, которые встречаются как в зоне J_1 файла *Dreiser1*, так и в зоне J_1 файла *Dreiser2*.

Пересечение выполнялось с помощью стандартных функций MSExcel ЕСЛИ ОШИБКА, ВПР, ЛОЖЬ. Для каждого слова в зонах пересечения была подсчитана вероятностная величина *P* по формуле (4). При подсчётах все числа округлялись до семи знаков после десятичного разделителя. В табл. 3 представлены статистические данные и результаты пересечения.

Для вычисления расстояния между текстами использовались следующие параметры.

$A=J_1(Dr1)\cap J_1(Di)$ – слова, входящие в пересечения зон J_1 текстов *Dreiser1* и *Dickens*;

$B=J_1(Dr1)\cap J_1(Dr2)$; слова, входящие в пересечения зон J_1 текстов *Dreiser1* и *Dreiser2;*

$T(A)$ = 1906 – общее количество слов в зоне пересечения *A*;

$T(B)$ = 2341 – общее количество слов в зоне пересечения *B;*

$w_i(A) = w_j \in A$, A={w1….w1906} – некоторое конкретное слово, являющееся элементом зоны пересечения *A*;

$w_i(B) = w_j \in B$, B={w1….w2341} – некоторое конкретное слово, являющееся элементом зоны пересечения *B*;

$Pw_i(Di)$ – вероятностная величина слова w_j в зоне J_1 текста *Dickens*;

$Pw_i(Dr2)$ – вероятностная величина слова w_j в зоне J_1 текста *Dreiser2*;

$Pw_i(Dr1)$ – вероятностная величина слова w_j в зоне J_1 текста *Dreiser1*;

$R1w_i(A)=\left|Pw_i(Dr1)-Pw_i(Di)\right|$ – разница между вероятностными величинами слова из зоны пересечения *A* в тексте *Dreiser1* и этого же слова в тексте *Dickens*;

$R2w_i(B)=\left|Pw_i(Dr1)-Pw_i(B)\right|$ – разница между вероятностными величинами слова из зоны пересечения *A* в тексте *Dreiser1* и этого же слова в тексте*Dreiser2*;

$K1=\sum_{w_{i=1}}^{190_6} R1w_i(A)=0,0816185$ – сумма разниц вероятностных величин слов, входящих в зону пересечения *A*;

$K2=\sum_{w_{i=1}}^{234_1} R2w_i(B)=0,0632247$ – сумма разниц вероятностных величин слов, входящих в зону пересечения *B*;

$L1=K1/T(A)$ = 0,0816185 / 1906 = 0,0000428. Этот параметр представляет собой среднюю сумму разниц и указывает на расстояние между *Dreiser1* и *Dickens*;

$L2=K2/T(B)$ = 0,0632247 / 2341 = 0,0000270. Этот параметр представляет собой среднюю сумму разниц и указывает на расстояние между *Dreiser1* и *Dreiser2*.

Таким образом, расстояние между работами одного автора *Dreiser1* и *Dreiser2* в 1,59 раза (0,0000428 / 0,0000270 = 1,59) меньше, чем между работами разных авторов *Dreiser1* и *Dickens*.

Таблица 3

Статистические данные трёх текстовых файлов

Текст	Общее количество токенов	Количество уникальных слов	Количество слов в J_1	Количество слов в зоне пересечения с Dreiser1	Примеры слов в зоне пересечения	*P*
Dickens	991471	22809	3234	1906	state white situation	0.0003681 0.0002017 0.0000403
Dreiser2	722838	21303	3143	2341	state white situation	0.0002822 0.0002615 0.0001881

В настоящей статье нами предложен оригинальный метод зонально-корреляционного анализа, предполагающий сопоставление распределения слов в зонах J_1 двух или более текстов. Этот метод предусматривает последовательное выполнение следующих процедур.

1. Разбивка словарного состава текстов на зоны J_0, J_1, J_2 по методике зонального анализа на основе разработанной ранее Y-интерпретации закона Брэдфорда. В соответствии с этой интерпретацией, указанные три зоны рассматриваются как подмножества, выделяемые из множества слов текста на основе их статистических параметров (частотностей). Зона J_0 содержит стоп-слова, с одинаково высокой частотностью встречающиеся в различных текстах; зона J_1 включает слова, представляющие основное содержание текста; зона J_2 состоит из редко используемых и встречающихся слов.

2. Выполнение пересечения зон J_1 сопоставляемых текстов. В результате пересечения создаётся зона, в которую входят слова, встречающиеся в соответствующих зонах сопоставляемых текстов.

3. Анализ (корреляция) параметров слов в зоне пресечения. В настоящей статье в качестве основного параметра использовались вероятностные величины. Для каждого слова, входящего в зону пересечения, подсчитывалась вероятностная величина в каждом из сопоставляемых текстов, далее высчитывалась разница по модулю между двумя величинами, а расстояние между текстами определялось по сумме разниц.

Зонально-корреляционный анализ может применяться для решения проблемы автоматической классификации и авторской атрибуции текстов. Обычно в таких случаях используются десятки, а то и сотни параметров. В одной из предыдущих работ [11] нами было выделено сорок пять параметров различных видов с целью распознавания художественных и нехудожественных (научных и газетных) текстов. В работе М. Сантини [12] описываются несколько сотен разделённых на три группы параметров, которые применяются для распознавания жанров веб-документов в процессе информационного поиска.

Проведённый зонально-корреляционный анализ позволяет получать адекватные результаты при использовании одного параметра – вероятностных величин слов. Это не исключает применения более сложных метрик и анализа других единиц текста (основ слов, n-gram), что является предметом дальнейших исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Яцко В.А. Компьютерная лингвистика или лингвистическая информатика? // Научно-техническая информация. Сер.2. – 2014. – № 5. – С.1-10.
2. Piantadosi S.T. Zipf 's word frequency law in natural language: a critical review and future directions. – 2014. – URL: http://colala.bcs.rochester.edu/papers/ piantadosi2014zipfs.pdf
3. West M. The mystery of Zipf. – 2008. – URL: http://plus.maths.org/content/mystery-zipf
4. Sorell J. Zipf's law and vocabulary //The encyclopedia of applied linguistics / ed C. A. Chapelle). – Oxford: Blackwell, 2012. – URL: http://www.academia.edu/550703/Zipfs_Law_and_Vocabulary
5. Большакова Е.И., Клышинский Э.С., Ландэ Д.В. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. – М.: МИЭМ, 2011. – 272 с.
6. Могилев А.В., Листрова Л.В. Информация и информационные процессы. Социальная информатика. – Санкт-Петербург: БХВ-Петербург, 2006. – 240 с.
7. Яцко В.А. Интерпретация закона Брэдфорда в терминах геометрической прогрессии // Научно-техническая информация. Сер. 2. – 2012. – № 4. – С. 19-23.
8. Anthony L. AntConc3.1.3. – 2012. URL: http://www.antlab.sci.waseda.ac.jp/antconc_index.html
9. Яцко В.А. Метод зонального анализа данных // В мире научных открытий. – 2013. – № 6.1. – С. 166-182.
10. Hadi W.M., Thabtah M., Abdel-jaber H. A comparative study using vector space model with k-nearest neighbor on text categorization data // Proceedings of the World congress on engineering / eds S. I. Ao, L. Gelman, D.Hukins, A. Hunter, A. M. Korsunsky. – London, 2007. – Vol. 1. – P. 296-300. – URL: http://www.iaeng.org/publication/ WCE2007/WCE2007_pp296-300.pdf
11. Яцко В.А., Стариков М. С., Бутаков А. В. Автоматическое распознавание жанра и адаптивное реферирование текста // Научно-техническая информация. Сер.2. – 2010. – № 5. – С. 9-18.
12. Santini M. Description of 3 feature sets for automatic identification of genres in web pages. – 2005-2006. – URL: http://www.nltg.brighton.ac.uk/home/ Marina.Santini/three_feature_sets.pdf

Материал поступил в редакцию 06.06.14.

Сведения об авторе

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор Хакасского государственного университета им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2’367.7

Е.А. Валова

Синтаксические свойства энклитической частицы *же* в диахроническом аспекте: корпусное исследование

На основе данных Национального корпуса русского языка показано, как на русских энклитиках сказался распад древнерусской системы расстановки энклитик, основанной на законе Ваккернагеля. На примере клитики «же» рассмотрено, какие позиции в предложении может занимать эта частица, а какие для нее невозможны. Продемонстрировано, какое влияние на выбор позиции частицы «же» оказывают особенности контекста и как менялось ее употребление с течением времени.

Ключевые слова: *же, порядок слов, энклитики, закон Ваккернагеля, русский язык, корпусное исследование*

Настоящая статья посвящена диахроническому исследованию синтаксических свойств русской энклитической частицы *же*, основанному на данных Национального корпуса русского языка. В рамках исследования мы ставили перед собой задачу ответить на следующий вопрос: как сказался на поведении частицы *же* распад древнерусской системы расстановки энклитик, основанной на законе Ваккернагеля, в соответствии с которым они имели тенденцию располагаться после первого полноударного слова фразы.

На примере энклитики *же* мы рассмотрим, какие позиции в предложении может занимать эта частица, а какие для нее невозможны. Будет показано, какое влияние на выбор позиции для частицы *же* оказывают особенности контекста и как менялось ее употребление с течением времени. Для этого мы представим диахроническую статистику употребления клитики *же*, основанную на материалах Корпуса, а также обрисуем современную картину ее употребления.

Клитиками называются слова, не имеющие собственного ударения или утрачивающие его и подчиняющиеся акцентуации предшествующего или последующего слова («носитель клитики», «опорное слово»). В первом случае акцентно несамостоятельное слово называется энклитикой, во втором – проклитикой. Вместе с опорным словом клитики образуют фонетический комплекс с единым ударением. Согласно [1, с. 28], клитики представляют собой «переходные феномены», «промежуточный класс единиц», которые нельзя отнести ни к словоформам, ни к морфемам: клитики менее автономны, чем словоформы (не могут составлять высказывания), но более самостоятельны, нежели морфемы (обладают свойством отделимости, зачастую также переместимы).

Утрата ударения некоторыми частицами и местоимениями, называемая энклизой, встречалась еще в индоевропейском праязыке. Подобное явление позднее имело место и в славянских языках, в том числе в русском, и нередко приводило к образованию сложных слов (ср. энклиза имен: *покамест* из *по ка места*).

Таким образом, среди энклитик можно выделить как древние (которые перешли в разряд акцентно несамостоятельных слов много веков назад, ср. *же*, *ли*), так и относительно новые (*бы*).

Закон Я. Ваккернагеля, сформулированный им в 1892 г. [2], «в своем наиболее общем виде гласит, что в древних индоевропейских языках энклитики располагаются так, что они составляют конечную часть первого фонетического слова («тактовой группы») фразы» [3, с. 45]. Иными словами, энклитики следуют после первого полноударного слова фразы – независимо от того, к какому грамматическому классу оно относится.

В монографии [4] специально оговаривается тот факт, что во многих случаях имеет смысл говорить о менее строгом варианте закона Ваккернагеля, согласно которому энклитики могут располагаться в предложении не только после первого полноударного слова, но также и после первой синтаксической составляющей. При этом некоторые языки придерживаются единой стратегии расположения ваккернагелевских энклитик, а в других возможно варьирование.

Одной из наиболее значимых работ по данной проблеме является содержащая детальное описание

системы энклитик древнерусского языка книга А.А. Зализняка [5], в основу которой легло изучение древних письменных памятников – берестяных грамот, наиболее точно отражающих характерные черты живой речи раннедревнерусского периода (XI – начало XIII вв.). По словам автора, ранние грамоты (XI – 1-я треть XIII в.) отражают древнейшее состояние системы энклитик, тогда как поздние (2-я треть XIII – XV в.) ближе к современному состоянию языка [5, с. 23].

В книге говорится, что к числу энклитик относились частицы *же, ли, бо, ти* (древнейшие энклитики), местоимения в дательном и винительном падежах (*ми, ти, си, ны, вы, на, ва,* и др.) и «относительно молодые» энклитики – связки (*есмь, есмы, еси, есте, еста, есть, суть* и др.), *бы*. Древнейшие энклитики, по словам автора, являются сильными: они «совершенно устойчивы в своих энклитических свойствах», тогда как у молодых встречается и акцентно самостоятельный вариант.

Древнерусские энклитики, относящиеся к сказуемому, подчинялись закону Ваккернагеля, те же, которые относятся по смыслу к какому-либо другому слову, располагались непосредственно после данного слова. А. Зализняк называет их «фразовыми», или «ваккернагелевскими» [5, с. 25], и «локальными» («неваккернагелевскими») соответственно и концентрирует свое внимание на первых.

Древнерусское *же* Зализняк называет фразовой энклитикой первого ранга [5, с. 28]: в блоке энклитик *же*, согласно рангу, занимало первое место. Именно данная энклитика лучше всего сохранила свои «ваккернагелевские» свойства [5, с. 48]: она располагалась после первого полноударного слова фразы (независимо от наличия или отсутствия проклитики перед ним). Перенос *же* вправо (как и перенос любой другой энклитики, а также их кластера) мог объясняться наличием во фразе ритмико-синтаксического барьера: начальная часть клаузы могла быть отчленена (пауза в произношении), и тогда закон Ваккернагеля действовал в той части клаузы, которая следует за отчлененной. Подобную точку «условного начала фразы» (после отчлененной части) А. Зализняк и называет ритмико-синтаксическим барьером [там же]. Таких барьеров во фразе могло быть несколько.

Схожее явление можно наблюдать и в современном языке. Так, например, фраза *Вчера Петя же этого еще не знал* произносится с маленькой остановкой после *вчера* или, по крайней мере, «с некоторым протяжением конечного *а* в этом слове» [там же,], а на уровне смысла слово *вчера* кажется выделенным. Поэтому, как считает автор, естественно предполагать здесь наличие ритмико-синтаксического барьера после первого слова, которое оказывается, таким образом, выведенным из сферы действия закона Ваккернагеля. С точки зрения А. Зализняка, этим и объясняется тот факт, что противительное *же* в некоторых фразах может «уходить вправо». Кроме того, автор отмечает, что в современном русском языке допускается постановка *же* после нескольких полноударных слов, «если они образуют смысловое единство» [там же]: *Железную дорогу же ему не оплатят*. Ниже будет рассмотрен сравнительно узкий класс подобных предложений.

Итак, в древнерусском языке фразовые энклитики (в том числе рассматриваемое нами *же*), а также блоки энклитик могли располагаться на «главном ваккернагелевском месте» [5, с. 50] (непосредственно после первого полноударного слова фразы при отсутствии ритмико-синтаксических барьеров перед ним), либо на «дополнительном» (которое определялось одним из возможных для данной клаузы барьеров).

Зализняк отмечает тот факт, что в смысловом отношении часть древнерусских фразовых энклитик тяготела не к сказуемому, а к тому члену предложения, который был «поставлен в начало фразы (как правило, это тема)» [5, с. 67]. Данное явление обусловлено тем, что сильные энклитики чаще других отрывались от блока энклитик, оказываясь левее ритмико-синтаксического барьера, и впоследствии могли образовывать устойчивые сочетания со словами, которые часто выступали в начальной позиции во фразе, и даже образовывали вместе с ними единые слова (ср. *неужели* из *не+у+же+ли*). Таким образом, мы можем говорить о некоторой эволюции энклитик, которые ранее относились к фразе в целом и подчинялись закону Ваккернагеля: постепенно фразы, их содержащие, были переосмыслены, и с точки зрения современного русского языка эти энклитики «должны рассматриваться просто как подчиненные предшествующему слову» [там же]. Зализняк предполагает, что в подобных случаях происходило «смысловое переподчинение энклитики» [5, с. 68]. При этом порядок слов после изменения смысловых отношений обычно сохранялся, и закону Ваккернагеля подчинялись не только энклитики, которые относятся к сказуемому, но и те, которые данное свойство уже утратили (к числу последних, по словам автора, можно отнести современное противительное *же*). В большей степени это утверждение справедливо для сильных энклитик.

Нашей задачей было установить, насколько современная ситуация отличается от той, что мы видим в древнерусском языке. Мы задались целью изучить позиционные свойства русской энклитической частицы *же* на материале Национального корпуса русского языка. При этом основное внимание было направлено на описание диахронических изменений и современной языковой ситуации, а также на выявление позиционных особенностей данной энклитики.

Сначала мы обратились к материалам грамматик и словарей, в соответствии с которыми *же* может считаться союзом либо частицей [6–9]. Мы в нашем исследовании подобного разделения не придерживаемся и называем *же* энклитической частицей. Энклитике *же* словари приписывают следующие значения: противительное (*Вася засмеялся, Петя же заплакал*), усилительное (*Дайте же мне сказать!*), присоединительное (*Я дружу с ним с тех пор, как его знаю, знаю же я его с детства*), отождествительное (*Я живу все там же*). Отдельно в словарных статьях указываются устойчивые выражения: *все же*, *как же* и др. Их мы исключили из рассмотрения, равно как и сочетания *же* с вопросительными словами.

Исследование проводилось на весьма узком классе примеров с частицей *же* и именами собственными. Нас интересовали те случаи, когда энклитика следует после двух имен собственных (например, имени и отчества) или же располагается между ними. В первом случае, как нам кажется, можно говорить о «нестрогом» варианте позиции Ваккернагеля: в качестве первого полноударного элемента предложения выступают два слова, синтаксически связываемых в единое целое.

Мы отобрали все примеры с интересующими нас комбинациями из Национального корпуса русского языка. Стоит оговорить, что для дальнейшего определения значений *же* необходимо было также учитывать «левый контекст», поэтому отобранные нами примеры зачастую выходят за рамки одного предложения.

(1) — На Шестнадцатую! — честно призналась я. — Так ***Павел Михайлович же*** *устал!* [Татьяна Соломатина. Мой одесский язык (2011)]

(2) Девки знали, какой юмор их ждет от Артюхи Колотушкина, воспринимали его трудно, плевались, покидали поляну. ***Артюха же Колотушкин*** *вдохновлялся пуще прежнего...* [Виктор Астафьев. Обертон (1995–1996)]

В поэзии представлена лишь вторая комбинация, в которой частица *же* следует после первого имени собственного.

Следующим шагом необходимо было исключить из рассмотрения нерелевантные примеры, в том числе примеры с дублированием и последовательностями клитик.

Кроме того, мы не включали в выборку примеры с именами типа *Эль Греко*, *Ван Гог*, поскольку их первый элемент представляет собой не полноударное слово, а проклитику.

Отметим, что в книге А.А. Зализняка отдельно упоминается «локальное», неваккернагелевское *же*, т.е. *же*, относящееся к конкретному слову: *тамо же*, *а то далъ Иванъ же* ‘а это дал тоже Иван’ [5, с. 29]. В современном языке такое *же* представлено главным образом в отождествительном значении.

Интересно, что найденные нами примеры подобного употребления *же* выглядят несколько иначе, нежели те, что приведены в [5] и словарях. Принципиальным отличием является наличие в предложении, содержащем подобное *же*, лексического повтора:

(3) — Меня не было, зато был Дмитрий Федорович, и я слышал это своими ушами от ***Дмитрия же Федоровича****...* [Ф.М. Достоевский. Братья Карамазовы (1880)]

(4) ...и Николай Александрович, разбухая, приобретал печать Адама Кадмона, не отличавшегося от ***Николая же Александровича****...*[Андрей Белый. Между двух революций (1934)]

(5) Так вот, тетя Дуся, считая, что дружба со мной идет на пользу гениальному Самсон Есеичу (...), а потому, уважая меня, через посредство ***Самсона же Есеича*** *дала мне ключ от своей комнатенки.* [Асар Эппель. Бутерброды с красной икрой (1990–2000)]

Нам встретилось шестнадцать подобных примеров. Мы не включали их в общую статистику, поскольку они представляют собой случаи «неваккернагелевского» употребления энклитики *же*.

В результате из основного корпуса были отобраны 507 примеров, которые затем были разделены на группы по временным интервалам в пятьдесят лет, после чего для каждого интервала подсчитывалось соотношение примеров с энклитикой *же* на втором и на третьем месте (табл. 1).

Изначально предполагалось, что для *же* в целом и для некоторых его значений в частности на протяжении последних 300 лет подчинение закону Ваккернагеля в его сильной формулировке постепенно становится менее обязательным. Эта гипотеза в ходе исследования подтвердилась. Как видно на графике (рис. 1), число примеров с *же* на третьем месте с течением времени несколько возрастает.

На следующем графике (рис. 2) наглядно представлено процентное соотношение примеров с обеими комбинациями для каждого интервала.

Таблица 1

Временной интервал	Число примеров с И+И+же	Число примеров с И+же+И	Всего примеров	Доля примеров с И+И+же	Доля примеров с И+же+И
1701–1750	0	2	2	0%	100%
1751–1800	0	1	1	0%	100%
1801–1850	1	21	22	4,5%	95,5%
1851–1900	4	155	159	2,51%	97,4%
1901–1950	17	131	148	11,5%	88,5%
1951–2000	16	103	119	13,4%	86,5%
2001–н.вр.	19	37	56	33,9%	66,0%
Общее число примеров	57	450	507	11,2%	88,7%

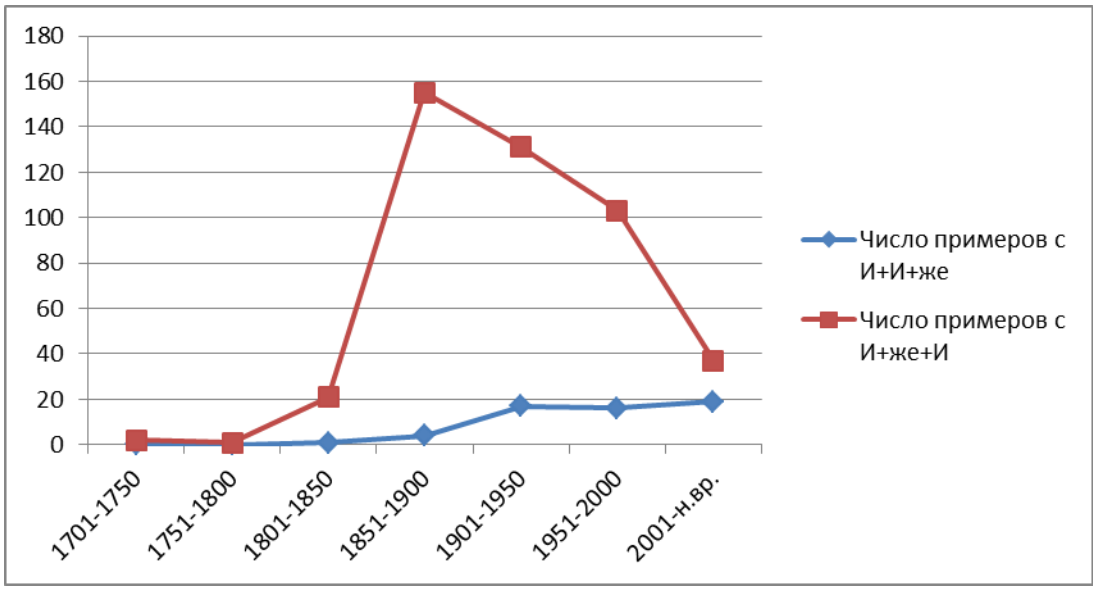

Рис. 1

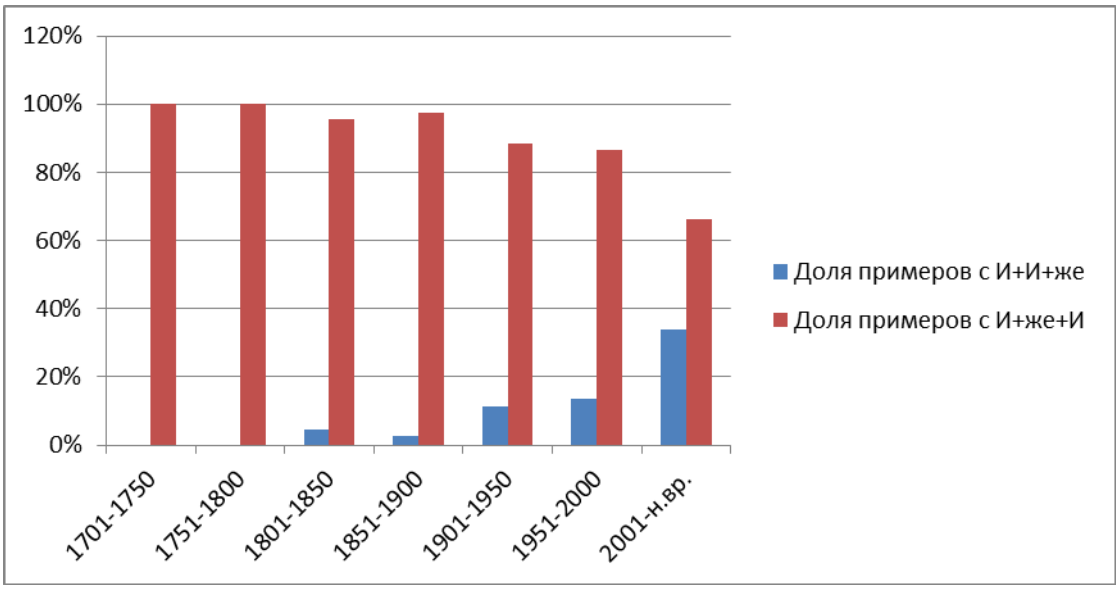

Рис. 2

Отдельно мы рассмотрели примеры с искомыми комбинациями из устного и газетного корпусов, которые, как естественно предполагать, наиболее наглядно отражают современную языковую ситуацию.

В устном корпусе нам встретилось всего пять примеров с *же* после имен собственных (все они относятся к концу XX – началу XXI вв.), из них лишь один пример – с *же* на третьем месте:

(6) [Бобчинский, Даниил Спиваковский, муж, 36, 1969] Нам сидеть никак нельзя. [Добчинский, Иван Агапов, муж, 40, 1965] У ***Петра Иваныча же*** *геморрой*. [Павел Лунгин и др. Дело о «Мертвых душах» / к/ф (2005)]

В оставшихся четырех примерах частица *же* подчиняется «строгому» закону Ваккернагеля и располагается после первого имени собственного.

Особый интерес для нашего исследования представляют данные газетного корпуса. В нем комбинация с *же* между двумя именами собственными встретилась 34 раза, тогда как случаев «нарушения» закона Ваккернагеля в два с половиной раза больше – 84. Отметим, что все примеры относятся к самому позднему временному интервалу. Таким образом, логично считать, что подобное соотношение искомых комбинаций подтверждает нашу гипотезу об ослаблении строго ваккернагелевского характера *же*.

Следующей нашей задачей было определить, какие факторы влияют на выбор позиции частицы *же* в предложении. В первую очередь было решено проверить, как соотносится позиция энклитики *же* с ее значением, поскольку изначально предполагалось, что в разных значениях частица *же* может обладать разными позиционными свойствами. Для этого мы рассмотрели отобранные примеры и определили значение *же* для каждого из них. В большинстве примеров (450) *же* представлено в противительном значении, значительно меньше предложений с усилительным *же* (55). В этой группе ввиду небольшого количества мы объединили примеры с собственно усилительным *же*, а также те, в которых клауза, включающая *же*, содержит уточнение или пояснение к сказанному ранее. Наиболее редко встречающееся значение *же*, присоединительное, представлено в двух примерах.

Противительное и усилительное *же* проявляют тенденцию к смещению вправо, тогда как присоединительное *же* в обоих примерах занимает позицию Ваккернагеля (табл.2).

В трех первых интервалах в примерах с энклитикой на третьем месте представлено только противительное *же*. В последующих интервалах его доля от общего числа подобных комбинаций колеблется от 75% до 84%, что объясняется общей частотностью данного значения энклитики.

В газетном же корпусе усилительное *же* оказывается в третьей позиции в 16,7% случаев (1 пример из 6), противительное – в 74% случаев (83 примера из 112).

В устном корпусе в единственном примере с «нарушением» строгого закона Ваккернагеля *же* имеет усилительное значение. В оставшихся четырех случаях на втором месте оказывается усилительное (3 примера) и противительное *же* (1 пример).

Же в присоединительном значении в устном и газетном корпусах не встречается.

Кроме того, *же* имеет тенденцию располагаться после комплекса иностранных имен собственных:

(7) ***Лион Фейхтвангер же*** *лично посетил этот процесс…* [Владимир Паперный. Вера и правда: Андре Жид и Лион Фейхтвангер в Москве (2003) // «Неприкосновенный запас», 2003.07.14]

Таких примеров в основном корпусе всего семь, все они относятся к XX–XXI вв.

В газетном корпусе нам встретилось двенадцать примеров с *же* после иностранных имен собственных.

Одним из факторов, блокирующих смещение энклитики *же* вправо относительно ваккернагелевской позиции, является наличие в предложении отрицания. В основном корпусе лишь в одном из двадцати двух примеров с отрицанием *же* располагается не на втором, а на третьем месте:

(8) "Автором года" стала Дарья Донцова. Ну не ***Джоан Ролинг же*** *её давать…*[Юлия Рахаева. Покупатель всегда прав. «Библио-Глобус» как зеркало книжных предпочтений (2003) // «Известия», 2003.02.11]

Отметим, что в газетном и устном корпусах во всех (5 и 2 соответственно) примерах с отрицанием *же* занимает в предложении вторую позицию.

В примерах с обращением *же* также занимает строгую позицию Ваккернагеля (всего 5 примеров в основном корпусе):

(9) Но, ***Алексей же Степанович****! — восклицал Коля...* [Максим Горький. Городок Окуров (1909)]

Отдельно стоит оговорить особые случаи, в которых позиция энклитики *же* определяется не общими правилами, а контекстом. Мы имеем в виду в первую очередь случаи, когда *же* ставится на второе место строго после выделяемого слова группы в целях различения двух людей с одинаковыми отчествами (6 примеров в основном корпусе, 1 – в газетном):

(10) Так Анастас Иванович уцелел при трех генсеках. ***Андрей же Иванович*** *при шести царствиях был у трона!* [А. Шубин. Путь к благополучию (2000)]

Итак, опираясь на изученные примеры, мы можем утверждать, что в усилительном и противительном значении *же* может встречаться в нестрогой позиции Ваккернагеля (после второго имени собственного), тогда как для присоединительного *же* единственно допустимой оказывается вторая позиция. Предварительное изучение других примеров с *же* показывает, что в сочетании с группой вида "прилагательное + существительное" частица *же* также может смещаться вправо относительно строгой позиции Ваккернагеля.

Кроме того, стоит отметить, что в газетном корпусе отклонения от строгого закона Ваккернагеля встречаются значительно чаще, нежели в основном корпусе.

Таким образом, проведенное исследование позволяет прийти к выводу, что со временем энклитическая частица *же* все чаще встречается в неваккернагелевской позиции, что объясняется общим разрушением древнерусской системы расположения энклитик, основанной на законе Ваккернагеля.

Таблица 2

Значение энклитики *же*	Число примеров с *же* на втором месте	Число примеров с *же* на третьем месте	Всего примеров
Противительное	401 (89,1%)	49 (10,9%)	450
Усилительное	47 (85,4%)	8 (14,6%)	55
Присоединительное	2 (100%)	0 (0%)	2

СПИСОК ЛИТЕРАТУРЫ

1. Плунгян В.А. Общая морфология. Введение в проблематику. – М.: Едиториал УРСС, 2003.
2. Wackernagel J. Über ein Gesetz der indogermanischen Wortstellung // Indogermanische Forschungen, Strassburg. – 1892. – Bd. I. – P. 333–434.
3. Зализняк А.А. «Слово о полку Игореве»: взгляд лингвиста. – М.: Языки славянской культуры, 2004.
4. Spencer A., Luis Ana R. Clitics: an introduction. – Cambridge, 2012.
5. Зализняк А.А. Древнерусские энклитики. – М.: Языки славянских культур, 2008.
6. Ушаков Д.Н. Толковый словарь русского языка. – М., 1948.
7. Ожегов С.И. Словарь русского языка. – М., 1989.
8. Даль В.И. Толковый словарь живого великорусского языка. – М., 1956.
9. Русская грамматика, 1980. – URL: http://rusgram.narod.ru/.

Материал поступил в редакцию 13.07.14.

Сведения об авторе

ВАЛОВА Евдокия Алексеевна – аспирант Национального исследовательского университета «Высшая школа экономики», Москва
e-mail: valova.dunya@gmail.com