ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 1** **Москва 2017**

ОБЩИЙ РАЗДЕЛ

УДК 002(091)

Е.А. Плешкевич

От документации к неодокументации

Анализируются материалы сборника трудов Академии документации (Тромсе, Норвегия), посвященного ее двадцатилетнему юбилею: история становления неодокументации, основные положения и перспективы ее развития.

Ключевые слова: *документ, документация, неодокументация, теория документа, Академия документации*

В 2016 г. исполнилось двадцать лет с момента создания Академии документации (*The Document Academy, DOCAM*), образованной на базе арктического университета Норвегии в Тромсе. Академия представляет собой международное общественное объединение ученых из разных стран. История ее создания связана с двумя событиями. Во-первых, с принятием 9 июня 1989 г. в Норвегии закона о создании юридического депозитария общедоступных документов (*The Norwegian Act of Legal Deposit of Generally Available Documents*). Согласно этому акту все опубликованные в Норвегии бумажные и печатные документы, фотографии, фильмы, материалы звукозаписи и цифровые offline и online публикации должны быть официально переданы на хранение в Национальную библиотеку Норвегии. Второе событие вытекало непосредственно из первого, и было связано с подготовкой специалистов в области документации. Оно включило разработку соответствующей образовательной программы и активизацию научных исследований. Образовательная программа получила название «Документационные исследова-

ния» (*Documentation Studies, Dovkit*) [1]. Ее автором стал научный коллектив университета в Тромсе под руководством Нильса Виндфельда Лунда (***Niels Windfeld Lund).*** Возвращение к идее документации было обусловлено, по мнению Лунда, несомненным достоинством документа как понятия, которое состоит в способности вместить в себя такие формы представления как изображение, письменные тексты, видео, танец, музыка и архитектура. Более того, это должно было подчеркнуть связь национальной библиотеки Норвегии, приступившей к собору всех видов и типов документов, с идеей всемирного дворца знаний, Мунданеума (*Mundaneum*), предложенной и частично реализованной Полем Отле.

Разработку новой программы поддержали такие известные ученые как Майкл Бакленд *(Michael Buckland)* и У.Б. Рейворд (*Warden Boyd Rayward*), однако в самом библиотечном сообществе Норвегии к этой программе отнеслись настороженно**.** Об этом свидетельствует тот факт, что программа была введена **только в университете Тромсе.** Занятия по ней начались в 1996 г. В 2013 г. образовательная программа была расширена и стала называться «*Media and Documentation Studies*».

Разработка этой программы стимулировала научные исследования в области документации и привела к созданию в Норвегии в 1996 г. Академии документации, первым руководителем которой стал Н. Лунд. Напомним, что документация как практическая деятельность и научная дисциплина возникла в первой половине XX столетия. Ее основателями выступили П. Отле, С. Брэдфорд, С. Брие. В связи с развитием в 1960–1970-х гг., с одной стороны, вычислительной техники и началом ее использования в научно-технической библиографии, а с другой стороны, – с разработкой теории информации произошел отказ от документации в пользу информатики и научно-информационной деятельности. Начиная с 2003 г. Академией стали проводиться ежегодные научные конференции сначала в США в Школе информации университета в Беркли, позже в других странах – Канаде, Швеции, Норвегии и Австралии. В 2016 г. конференция состоялась в начале сентября в университете Северного Техаса в Дентоне. С 2014 г. начинают издаваться труды Академии. В 2016 г. вышел третий юбилейный том [1], посвященный неодокументации. Мы остановимся на анализе тех статей этого тома, которые раскрывают основные положения новой концепции документации, которая получила приставку «нео».

Содержание учебной программы подробно раскрыто в статье Н. Лунда «Как это все начиналась: 1996, первый год *Dovkit*» (*How It All Started: 1996, the First Year of Dovkit)*. Программа предназначалась для обучения студентов старших курсов, а также магистрантов и докторантов (PhD). Обучение по программе подразделялось на базовый и продвинутый уровни. На освоение базового уровня отводились первые два семестра; третий семестр отводился на изучение продвинутого уровня, после которого магистранты в течение 4 и 5 семестров изучали библиотековедение (*Library Studies*); 6 и 7 семестр отводился для изучения дисциплин по выбору. Таким образом, можно сказать, что программа носит предварительный характер: структурно программа включала гуманитарный, социальный и естественно-технический компоненты, которые, в свою очередь, подразделяются на четыре образовательных блока или курса. Первые три из них посвящены изучению производства документов, их применению, распространению (рассеиванию) и поиску, четвертый носит обобщающий характер, сводя полученные ранее знания в единую систему. Содержательно она включает такие темы, как понятие документа (история документальной терминологии и документации как научной дисциплины), производство документов, документы на базе перспективных информационных технологий, организация науки, информационное право, компьютерные технологии, документационные системы и документальные институты. Проектная деятельность студентов ориентирована на освоение различных документационных процессов и технологий в сфере библиотечного, архивного и музейного дела. Анализируя перспективы дальнейшего развития программы *Dovkit*, Н. Лунд прогнозирует усиление медиа составляющей, что и послужило основанием принятых изменений в ее названии.

Вопросам методологии документационных исследований посвящена статья американского библиотековеда и документалиста М. Бакленда[2] «Физический, ментальный и социальный аспекты документов» (*The Physical, Mental and Social Dimensions of Documents*). В одной из статей, написанной ранее [2], была высказана гипотеза о том, что феномен документа может быть раскрыт посредством одновременного рассмотрения его физического, ментального и социального аспектов. В своей статье Бакленд развивает эту гипотезу. Так, физический аспект, по его мнению, связан с тем, что тексты или произведения могут существовать в качестве документов только как физические явления. Физическая составляющая обеспечивает существование документов во времени и пространстве, а технологии их производства, включая письменность, печать и телекоммуникации, направлены на непрерывное сокращение ограничений, налагаемых временем и пространством. Однако, для того чтобы физически созданные текст или произведение рассматривались в качестве документов, они должны восприниматься как доказательства или свидетельства. Бакленд связывает это восприятие или перцепцию[3] с ментальным аспектом документа. Отсюда следует, во-первых, субъективность и изменчивость восприятия объекта как документа; во-вторых,

[1] Некоторые отечественные исследователи переводят *Documentation Studies* как *документоведение*, однако в силу того, что в отечественной науке за этим термином закреплено понимание научной дисциплины, ориентированной на исследование управленческой документации и делопроизводства, то предлагаем переводить этот термин как *документационные исследования*.

[2] В нашей стране был ряд публикаций М. Бакленда по библиотечному делу [2], теории информатики и документации [3], а также краткие обзоры его научного творчества [4, 5].

[3] Перцепция (от лат. perceptio - представление, восприятие) – процесс непосредственного активного отражения когнитивной сферой человека внешних и внутренних предметов (объектов), ситуаций, событий, явлений и т.п.

особость познания документного восприятия. Результаты этой перцепции, отмечает он, могут быть наблюдаемы, однако саму перцепцию документа ни наблюдать, ни измерять не представляется возможным. В итоге она предстает в виде некоего «черного ящика». По Бакленду, социальный аспект тесно примыкает к ментальному аспекту и отличается от него тем, что имеет интерсубъективный характер, связанный уже не с индивидуальной, а с коллективной перцепцией документа как составной частью общего культурного кода.

Важное место в этой методике занимает моделирование возможных комбинаций аспектов документов. Так, комбинация физического и социального аспектов проявляется в разделении труда, в том, что узкая специализация обусловливает необходимость политической, управленческой и экономической координации. На практике эта координация реализуется посредством документов. При этом документальные технологии, начиная с письменности и заканчивая компьютерами, способствуют разделению труда. Исходя из этого, Бакленд полагает, что так называемое современное «информационное общество» могло быть более адекватно описано как «документальное общество». Комбинация социального и физического аспектов в ряде случаев обеспечивает регламентацию ментального аспекта. Примером этого выступает информационная политика, направленная на государственное регламентирование вопросов интеллектуальной собственности, использования учебной литературы, конфиденциальности, технических стандартов и т.п. Комбинацию социального и ментального аспектов он видит как использование письменных текстов в процессах сохранения и передачи культурных ценностей и в воспитании. Касается Бакленд и явления релевантности, которое он связывает с ментальным аспектом документа.

Исследованию лингвистической структуры понятия *документ* и перспективам его дальнейшего развития посвящена статья американского профессора библиотечно-информационной науки из Индианского университета в Блумингтоне Рональда Дэя (*Ronald Day*) *Sense in Documentary Reference: Documentation, Literature, and the Post-documentary Perspective*. Предпринятый им лингвистический анализ заключается в выделении значения (*reference*) и смысла (*sense*) понятия документ и исследовании их взаимосвязи. Как известно, различие между смыслом понятия и его значением было предложено немецким философом Готлобом Фреге (1848–1925). Под значением им понималась предметная область, соотнесенная с ее названием; под смыслом – сведения об этом названии. Сегодня в логической семантике значение языкового выражения называют его денотатом, т.е. предметом или классом предметов, которые обозначаются этим выражением; смыслом – то мыслительное содержание, которое выражается и усваивается при понимании языкового выражения.

Придерживаясь этой логики, под значением понятия *документ* Дей предлагает понимать его репрезентацию (*representation*), т.е. представление выражаемой этим документом предметной области. Истоки этого подхода были заложены еще П. Отле, который предложил рассматривать документ как литературную графическую или скульптурную форму представления реальности [7, с.288]. Комментируя этот тезис, Дей отмечает что, содержание документов, т.е. информация, содержащаяся в них, представляет мир, вернее его научную картину. Документальный процесс П. Отле видит в виде схемы, согласно которой знания как представление о мире возникает в сознании ученого и отображается им в различных документальных материалах. Далее идет научная работа с документацией, которая рассматривается П. Отле как высшая форма научной деятельности. При этом она включает не простое эмпирическое или экспериментальное формирование представлений о мире, а выработку истинных утверждений о нем. Согласно Дею, факты принадлежат не эмпирическим событиям, а скорее документальным записям и теориям, представленным в таких редуцированных формах как рефераты, таксономические классификации и библиотечные шифры.

Развивая лингвистическую концепцию документации, Дей рассматривает эпистемологические представления Отле о документах «как репрезентативные утверждения о фактах» и об универсальной библиографии как «представлении о мире» в сочетании с положениями логического позитивизма Л. Витгенштейна (1889-1951), изложенного им в «Логико-философском трактате» (1921). Напомним, что по Витгенштейну мир есть все, что происходит, следовательно, он состоит не из предметов или вещей, а из совокупности всех фактов. При этом вещи или предметы это всего лишь составная часть факта. Синтезировав эти идеи, Дэй приходит к умозаключению, которое в упрощенном виде можно сформулировать следующим образом: если книга или документ в широком смысле есть форма представления и выражения фактов, а мир есть не что иное, как совокупность всех фактов в их логическом пространстве, то сам мир или, по крайней мере, его структуру, можно представить как совокупность всех книг-документов, объединенных в коллекциях, библиотеках и каталогах. Последние, как мы понимаем, призваны образовывать необходимые логические связи между фактами, представленными документами.

Рассматривая документацию в качестве инструмента построения картины мира, Дэй предлагает проанализировать документ с точки зрения положений логического позитивизма. Он обращается к анализу того, как значение термина *документ* согласуется с его смыслом (*sense in reference*). Дэй отмечает, что согласно традиции, заложенной Отле, знания и истину можно представить через так называемое «изобразительное» описание сущности предметов и вещей («*aboutness» of things*), иными словами – представление их смысла. Опираясь на эти положения, Дей поднимает вопрос о том, каким образом от представления смысла отдельных предметов и вещей можно перейти к представлению универсальной картины мира. По его мнению, это становится возможным на основе представления информации в виде библиографии. Однако библиографический язык метаданных гораздо беднее естественного языка, что приводит к существенным искажениям. Дей называет это явление разрывом между смыслом и значением. В связи с этим современное развитие документации

связывается Дэем с приближением потенциала и гибкости языка метаданных к естественному языку.

Внедрение в практику документации новых компьютерных технологий, по мнению Дэя, должно способствовать улучшению качества индексируемых терминов и тем самым – значительному повышению точности и эластичности процессов получения новых знаний. При этом внедрение новых технологий обеспечивает переход от монологичного принципа поиска информации к диалогичному, что, в конечном счете, способствует «дрейфу» содержания понятия *документ* в направлении коммуникации.

В заключение Дэй отмечает, что век документации, включая соответствующие технологии и социокультурные аспекты, уходит в прошлое, однако, если взглянуть на все эти процессы с точки зрения новых технологий и исследований, то он сохраняется, но на более высоком уровне инфраструктурной интегрированности и абстрактности. В нашу эпоху понятия «смысл», «содержание» и «значение» сохраняют свою важность и нуждаются в дальнейшем осмыслении. Однако эта задача все более усложняется. Ее решение он частично связывает с переформатированием ключевого вопроса о документе. Напомним, что прежнее поколение документалистов интересовал вопрос о том, что есть документ (*What is a document*?). Современное поколение неодокументалистов должен интересовать вопрос о том, как о документе надо мыслить и каким он может быть (*How can document think and be?*).

В статье «Документ: многозначное понятие» (*The Document: A Multiple Concept*) исследователь из университета в Тулузе Сабина Ру (Sabene Roux) рассматривает развитие теоретических представлений о документе во Франции в последние 30-40 лет как трансформацию содержания документа от простого к сложному или многозначному (*multiple*). Под простым содержанием она предлагает считать трактовку документа с точки зрения потребителя. Напомним, что С. Брие (*Suzanne Briet*) еще вначале 1950-х гг. заявляла, что документ существует как таковой, поскольку потребители нуждаются в доказывании или объяснении чего-либо. Такой взгляд на документ, отмечает Ру, был поддержан библиотечным сообществом, однако в социальных науках стал вырабатываться другой подход к определению документа. Начало этому подходу было положено французскими учеными – социологом и писателем Роббером Эскарпи (*Robert Escarpit*) и специалистом в области информатики Жаном Мериа (*Jean Meyriat*).

Современное развитие трактовки документа, по мнению Р. Эскарпи (1918-2000), было связано с переориентированием понимания от понятийного к концептуальному[4]. В контексте разрабатываемой информационной и коммуникационной теории документа Эскарпи обращает внимание на фундаментальное различие между событием и документом. Под событием он понимает факт, имеющий место в определенном пространстве и времени, который невозможно воспроизвести заново, записать и передать. В отличие от события в документе факты или данные зарегистрированы в физической форме и, следовательно, могут быть воспроизведены. Он сравнивает документ со следами прошлого доступными для чтения и считает, что документ является средством для получения знаний при условии, что следы прошлого, благодаря использованию письменности, остаются доступными для чтения. В итоге он предлагает рассматривать документ как визуальный или тактильный информационный объект, обладающий дуальной стабильностью во времени. Синхронная или внутренняя стабильность во времени заключается в независимости содержания информационного объекта от линейной последовательности событий и способности сопоставлять многозначные следы прошлого. Внешняя или глобальная стабильность во времени заключается в способности объекта быть сохраненным, перемещенным и воспроизведенным. Что касается многозначности концепта документа, то она определяется разнообразием целей его пользователей.

К интенциональности [5] и многозначности понятия *документ* обращается Жан Мериа (1921-2010). Выступая на первом конгрессе Французского общества информации и коммуникации в 1978 г. он заявил, что любой объект может стать документом и что таковым его делают пользователь и получатель сообщения. Желание потребителя получить информацию, отмечает Мериа, выступает необходимым компонентом объекта, который рассматривается в качестве документа, хотя мотивы создателя этого объекта могли быть иными. Согласно этой дефиниции режим использования документа определяет его статус. Анализируя взгляды Мериа, Ру приходит к выводу, что в основе его представлений о документе лежит «феномен желания обучить или информировать», что информация по желанию ее получателя активируется в форме документа, и, следовательно, сам документ может быть рассмотрен как форма интенциональности информации. Бесконечно возможное количество пользователей обуславливают разнообразие в определении документа.

Далее Ру подробно анализирует статью Ж. Мериа «Документ, документация, документология» (*Document, documentation, documentologia*) [9], в которой автор предложил определять документ как объект поддержки информации, который используется для ее передачи, обеспечивая устойчивость коммуникации. Структура документа включает две составляющие –

[4] Различия между понятием и концептом состоит, прежде всего, в интенсионале (т.е. содержании, наборе релевантных признаков). Релевантность или соответствие того или иного признака соответствующей предметной области, его акцентированность определяется: во-первых, уровнями объективированности знания, отвлечения от субъективных факторов (более высокой степенью для понятия и меньшей для концепта);во-вторых, сферой функционирования (более широкой, в идеале общечеловеческой для понятия и более узкой, личной либо групповой, социоэтнокультурной для концепта) [8].

[5] Напомним, что интенциональность *в феноменологии рассматривается как первичная смыслообразующая устремленность сознания к миру, смыслоформирующее отношение сознания к предмету, предметная интерпретация ощущений.*

материальную, в виде контейнера, обеспечивающего поддержку коммуникационного процесса, и информационную – в виде контента. Исходя из этого, объект может считаться документом в силу выполнения им функций поддержки или передачи информации. Такое разделение позволило Мериа выделить два типа документов: «документы по интенции» ("*documents by intention*") и «документы по атрибуции» ("*documents by attribution*"). В первом случае объект рассматривается в качестве документа, если его создание служит интенции, т.е. намерению дать старт коммуникационному процессу. Во втором случае объекты становятся документами, когда пользователь использует их для поиска и получения информации. Эта дифференциация, по мнению Ру, позволяет представить документ как объект, обладающий несколькими информационными функциями, при этом единичный объект может с успехом выступать в качестве нескольких документов.

Переход к цифровым технологиям придал дополнительный импульс дискуссии о природе документа во французском научном сообществе. Так, Мериа заметил, что документ как любой другой продукт человеческой деятельности, созданный в пространстве, возникает в случае, когда различные социальные и технико-социальные системы встречаются. Жан-Поль Мецджи (*Jean-Paul Metzger*) и Женевьева Лалич-Бойдин (*Geneviève Lallich-Boidin*) предложили отказаться от дуализма в определении электронного документа [6] и рассматривать его в качестве новой коммуникационной технологии. Группа французских исследователей Национального центра научных исследований, объединившихся вокруг междисциплинарного сетевого сообщества «Документ» (*Multidisciplinary network "Document" RTR-DOC*) предложила определить документ в качестве антропологического термина, обозначающего представления фрагмента истины за пределами хаоса, какофонии и забвения. Сфокусировав внимание на нематериальной природе электронного документа, данная группа исследователей предложила идею «редокументаризации» (*redocumentarization*) как документальной материализации нематериальной информации, циркулирующей в сети (*a documentary materialization of immaterial information circulating on network*). В этом случае, считают они, электронный документ может быть определен как объект созданный автором и через некоторое время воссозданный другими. Более того по их мнению меняются функции документа в электронно-цифровой среде. Так, профессор в области информации и коммуникации Жан-Мишель Салайн (*Jean-Michel Salaün*) предложил рассматривать документ в качестве атрибута замещения. Анализируя документ как элемент сложной системы, Мериа предпринял попытку согласования представлений о традиционном и электронном документе. Он предложил обобщающее определение документа согласно которому, во-первых, документ, передающий информацию, не является простым носителем, имеет собственную сущность, заключающуюся в его неразрывной связи с информацией; во-вторых, документ имеет автора. Его нельзя игнорировать, он стремится к общению, что отражается в назначении документа; в-третьих, автор существует не только для того, чтобы продуцировать документы, он социален, может играть разные социальные роли и накладывать на документ различные ограничения; в четвертых, любой документ встроен внутрь коммуникационной системы и создан со специфическими целями.

Особое внимание Ру уделяет дискуссии в отношении материальной и знаковой природы документа. Профессор в сфере информации и коммуникации университета в Тулузе Вивиан Кузинэ (*Vivian Couzinet*) полагает, что документ – это форма организации информации, при которой контент принимает форму на коммуникационном и материальном уровнях для обеспечения последующей циркуляции. В соответствии с представлениями профессора университета Шарля де Голля в Лилле Анетты Биге-Вербуж (*Annete Beguin-Verbrugge*) электронный или аналоговый (традиционный) документ по существу должен определяться на основе его базовой функции, т.е. как то, что мы используем в качестве доказательства, представления информации и демонстрации существования чего-либо. Каролина Курбье (*Caroline Courbières*) из Университета Тулузы в контексте семиотико-лингвистического подхода рассматривает документ как информационный объект, созданный с коммуникационными целями и содержащий знак и его материальное обеспечение, как часть сложного информационно-коммуникационного «устройства». Она предлагает определять документ как артефакт, документальный статус которого устанавливает получатель. По ее мнению, документ сливается со знаком в момент его идентификации как результат интерпретации. Социальная ценность документа зависит от статуса его автора и его использования.

В заключение Ру излагает собственное оригинальное видение природы документа и методологию ее познания. В основе ее взгляда на природу документа лежит идея ризомы (ризоматической модели)[7]. В качестве иллюстрации этой идеи она моделирует процесс, в котором отчет о путешествии (*report of travel*), изначально, являясь единичным интенциональным и первичным документом, содержащим сведения, полученные опытным путем, продуцирует различные вторичные документы, которые в ряде случаев могут обладать большей ценностью, чем сам отчет. Совокупность первичных и вторичных, инициативных и атрибутивных документов, имеющих

[6] Авторы использует термин цифровой (*digital*) документ, однако, согласно принятой в нашей науке терминологии, мы переводим его как электронный.

[7] Ризома это одно из ключевых понятий философии постструктурализма и постмодернизма, введенное Ж. Делёзом и Ф. Гваттари в 1975 г., чтобы описать какую-либо теорию и какое-либо исследование, которые допускают множественные неиерархичные (не упорядоченные в какую-либо иерархию) точки входа и выхода в представлении и интерпретации знания. Наглядным примером для ризомы выступает запутанная корневая система растения. Согласно Делёзу и Гваттари, у ризомы нельзя выделить ни начала, ни конца, ни центра, ни центрирующего принципа («генетической оси»), ни единого кода.

разное значение, образует соответствующую ризому, которая в социальном пространстве продуцирует информацию, знание и является источником для художественного творчества. Ризоматическая модель документа раскрывает его многозначность и позволяет исследовать его отдельные аспекты вне контекста их иерархических связей. По мнению Ру, та модель альтернативна рассмотрению документа в качестве непрерывной модели (*continuum model*) с жесткими иерархическими связями и однозначным толкованием. Историческим примером такой ризомы документа выступают судовой дневник Ч. Дарвина и созданные на его основе такие вторичные документы как статьи и монографии, посвященные теории происхождения видов.

Таким образом, следует отметить, что феномен неодокументации связан пусть с локальной, на уровне Норвежской национальной библиотеки, но все же актуализацией основной идеи документации – формирование документационно-знаниевой картины мира. Представленный сборник в целом и анализируемые статьи в частности раскрывают историю этого науковедческого «апгрейда» и основные теоретические положения такого явления, как неодокументации. Ознакомление с этим изданием расширяет наши представления о путях построения соответствующей картины мира и о феномене документа как инструмента этого построения. Во-первых, сборник имеет важное научное значение и интересен всем, кто интересуется данной проблематикой.

Во-вторых, рассматривая эристический потенциал неодокументации и перспективы инкорпорирования отдельных ее положений в отечественную науку об информации и документе, прежде всего, следует отметить науковедческую связь неодокументации с отечественной информатикой (научно-информационной деятельностью), занимающейся формированием научно-информационной картиной мира. В нашей стране разработкой этой проблемы занимается научный коллектив ВИНИТИ РАН.

Теперь, что касается понятия документа. Как мы уже отметили, оно используется в неодокументации как обобщающая абстракция, позволяющая включить в знаниевую картину мира различные объекты, содержащие информацию – знание. Это известный методологический прием. В качестве иллюстрации можно привести использование К. Марксом при описании экономического устройства капиталистической системы понятия «товара». Напомним, что он наделил эту абстракцию только одним свойством – способностью участвовать в свободном обмене на другие вещи, оставив за рамками рассмотрения такие ее свойства, как полезность, целенаправленность ее производства для продажи и т. д. Это абстрагирование позволило К. Марксу использовать ее для обозначения не только результатов товарного производства, но и, к примеру, рабочей силы. При этом надо понимать, что эта сила не является товаром в строгом онтологическом товароведческом смысле. Придерживаясь аналогичной логики, можно сказать, что документ в неодокументции не столько исследуется, сколько гносеологически конструируется под решение определенной задачи. И именно так его необходимо и рассматривать. Признание этого освобождает нас от идеи создания всеобщей науки о документе, будь то документология или что-то подобное. Вполне очевидно, что оценить эффективность такого гносеологического конструирования в полной мере можно лишь по результатам соответствующей документационно-знаниевой картины мира, которая к настоящему времени находится в процессе разработки. Что же до сих пор мешает ее созданию? Возможно, недостаточная проработанность самого концепта «документ». Решение этой задачи С. Ру видит в развитии его многозначности через создание ризоматической модели. Однако возможна и другая причина, связанная с появлением таких более фундаментальных абстракций как информационный (научно-информационный) объект и процесс. В любом случае дать однозначный ответ пока что не представляется возможным.

Методологические особенности неодокументации заключаются в ориентации на логический позитивизм, делающий акцент на знаковой природе документа. Напомним, что еще Отле, под документом предлагал понимать все, что графическими знаками изображает какой-либо факт или идею. Это важное направление, приближающее нас к пониманию природы документа, однако не вполне понятно, почему его рассмотрение должно быть ограничено позицией получателя или создателя информации. Во-первых, говоря о социальной природе документа, мы должны рассматривать документ, в том числе и с позиции общества как участника коммуникационных процессов. Во-вторых, нам представляется, что рассматривая документ в качестве знака, акцент необходимо сделать на исследовании его особых свойств или качеств, отличающих его от других знаков. По нашему мнению, с семиотической точки зрения особость документа как знака заключается в феномене институционального обеспечения соответствия значения знака, отраженного в его названии и содержании его смыслу, что мы трактуем как обеспечение семантической симметрии сообщения. Очевидно, что в условиях социального и экономического неравенства, столкновения интересов различных членов общества и других общественных противоречий, имеет место манипуляция информацией, передаваемой при помощи знаков, что ведет к искажению информационной картины мира. На практике это проявляется в феномене целенаправленного (в редких случаях случайного) искажения смысла посредством подлога, либо ограничения доступа вплоть до уничтожения или утраты. Следствием этого стала постепенная утрата доверия к письменному сообщению как соответствующему регулятору общественных отношений и отказ от использования переданной с его помощью информации. Разрешение этого информационного противоречия заключалось в разработке специальных механизмов обеспечения необходимого уровня соответствия значения и смысла.

В устных коммуникациях проблема соответствия значения смыслу разрешалась на основе личностного доверия между участниками процесса, а также при помощи института свидетелей. При переходе к письменным коммуникациям институт свидетелей, в силу

своей низкой мобильности и по ряду других причин, постепенно стал утрачивать возможности обеспечения необходимого уровня соответствия между значением и смыслом. Выход из сложившейся ситуации был найден через создание письменных знаков и их использование, по крайней мере, в тех сегментах информационно-коммуникационной деятельности, где без этого нельзя было обойтись. На практике эти письменные сообщения получили название как документы, а технология их создания как документирование. Таким образом, с точки зрения логического позитивизма, документы – это письменные и иные знаки, где необходимый уровень соответствия значения и смысла обеспечивается через такие социальные институты, как делопроизводство и нотариат, библиотечно-библиографические учреждения, архивы, службы НТИ и дата-центры. Аккумулирование в некоторых из этих информационных институтов письменных и иных сообщений создает возможность формирования соответствующей научно-информационной картины мира и как результат – развития науки как таковой. Письменные знаки, где отсутствует институциональная форма обеспечения доверия между значением и смыслом, мы предлагаем рассматривать как протодокументы.

СПИСОК ЛИТЕРАТУРЫ

1. Neo-documentation Around the World: Global Development // Proceedings from the Document academy. – 2016. – Vol. 3, Is. 1. – URL: http://ideaexchange.uakron.edu/docam/about.html
2. Olsen B.I., Lund N.W., Ellingsen G., Hartvigsen. Document theory for the design of socio-technical systems: a document model as ontology of human expression // Journal of documentation. – 2012. – Vol.1 (68). – P. 100-126.
3. Бакленд М. Модернизация библиотечного дела: манифест. Как привлечь внебюджетные средства, стать фандрайзером: Принципы ипрактики развития б-ки / пер. с англ. Виктория Стил, Стивне Д. Элдер. – М. : О.Г.И., 2001. – 268 с.
4. Бакленд М. Какого рода наукой может быть информатика? // Международный форум по информации. – 2012. – №2. – С. 3–9.
5. Плешкевич Е.А. Феноменологическая теория документа Майкла Бакленда: сущности и перспективы развития // Научно-техническая информация. Сер.1. – 2014. – №8. – С.35-40.
6. Лиховид Т.Ф. Майкл Бакленд: понятие информации в библиотечно-информационной сфере // Румянцевские чтения – 2013. Материалы международной научной конференции (16-17 апреля 2013 г.). Ч.1. – М. : РГБ, Изд-во «Пашков дом», 2013. – С. 360-365.
7. Отле П. Библиотека, библиография, документация :избр. тр. пионера информатики / пер. с англ. Р.С. Гиляревского и др. – М. : ФАИР-ПРЕСС, Пашков дом, 2004. – 350 с.
8. Сусов А.А., Сусов И.П. Размышления о концептах // Вісник Харківського національного університету ім. В.Н. Каразіна. № 726. Серія: Романо-германська філологія. Методика викладання іноземних мов. – Вип. 49. – Харків, 2006. – С. 14–20. – URL: http://homepages.tversu.ru/~ips/ASusov2006a.html
9. Meyriat Jean. Document, documentation, documentologia // Schèma et schèmatisation. – 2e trimester. – 1981, – № 14. – P. 51-63.

Материал поступил в редакцию 11.10.16.

Сведения об авторе

ПЛЕШКЕВИЧ Евгений Александрович – доктор педагогических наук, Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук, главный научный сотрудник (Новосибирск)
e-mail: eap1966eap@mail.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 025.4 : 004.89

И. М. Зацман, О. Ю. Инькова, В. А. Нуриев

Построение классификационных схем: методы и технологии экспертного формирования*

Рассматривается задача построения классификационной схемы логико-семантических отношений между частями предложений, предложениями и фрагментами текста вне зависимости от языка, на котором текст написан. Предлагаемая технология построения классификационной схемы включает две основные стадии: автоматизированное формирование перечня классификационных рубрик и создание схемы на основе сформированного перечня. Разработанный метод автоматизированного формирования рубрик позволяет создавать верифицируемые классификации в широком спектре предметных областей, использующих методы обработки текстов и других информационных объектов, в том числе в сфере научно-технической информации.

Ключевые слова: *верифицируемые классификации, автоматизированное формирование классификационных схем; корпусная лингвистика; логико-семантические отношения; информационная технология*

ВВЕДЕНИЕ

Построение универсальной классификации логико-семантических отношений (*далее* – ЛСО) – основная задача проекта «Логическая структура текста: контрастивный анализ способов выражения логико-семантических отношений в русском, французском и итальянском языках», который в настоящее время выполняется в Институте проблем информатики ФИЦ ИУ РАН. Об актуальности этой задачи свидетельствует факт разработки стандарта ISO по схеме и методике аннотирования логико-семантических отношений. В процессе разработки этого стандарта сопоставляются существующие методы и технологии построения классификаций ЛСО, анализируются общие черты и различия существующих классификационных схем, предлагаются их модификации [1, 2].

Применяемые в информатике и компьютерной лингвистике методы и технологии построения таксономий и классификационных схем можно условно разделить на две основные категории в зависимости от основной цели и задач их построения. В технологиях первой категории ставится задача семантического аннотирования экспертами информационных объектов, например, лингвоспецифичных единиц, глагольных конструкций или языковых реализаций логико-семантических отношений [3–8]. Такие технологии будем называть экспертными. Аннотирование выполняется лингвистами-экспертами в процессе семантического анализа текстов одного или нескольких корпусов [9–11]. Аннотированные информационные объекты используются, в частности, и как эталон для оценки качества программ обработки текстов на естественном языке [12].

В технологиях второй категории ставится задача автоматического, т.е. без участия человека, извлечения таксономий информационных объектов из текстов. С этой целью создаются программы для анализа текстовых фрагментов, в том числе на основе методик, правил или алгоритмов, полученных в результате использования технологий первой категории. Иначе, технологии второй категории, которые будем называть компьютерными, создаются, как правило, на основе методов извлечения, представления и организации конвенциональных и новых знаний, полученных в процессе семантического анализа текстов [13–15].

Каждая из двух категорий (экспертные и компьютерные технологии) может быть, в свою очередь, разделена на две подкатегории: одноязычную и многоязычную. Вторая подкатегория предполагает проведение контрастивного анализа параллельных текстов на двух или более языках с помощью экспертных или компьютерных технологий. Пример экспертной многоязычной технологии обработки параллельных текстов на пяти языках рассмотрен в работе [11], где анализируются коннекторы английского языка *although, because, also, if* и устанавливаемые

*Исследование выполнено в Федеральном исследовательском центре «Информатика и управление» РАН за счет гранта Российского научного фонда (проект № 16-18-10004).

ими ЛСО, а также способы их перевода (в том числе и нулевой эквивалент) на нидерландский, немецкий, французский и испанский языки.

В рамках предлагаемой типологии технология построения классификационной схемы ЛСО является экспертной многоязычной. Она состоит из двух основных стадий, которые могут повторяться:

- автоматизированное формирование (уточнение) перечня классификационных рубрик, которое выполняется экспертами – разработчиками схемы в процессе аннотирования показателей ЛСО в параллельных текстах на трех языках (русском, французском, итальянском);
- создание верифицируемой классификации ЛСО на основе сформированного (уточненного) перечня.

Использование в нашем проекте текстов параллельных корпусов на трех языках позволяет, с одной стороны, предложить универсальную классификацию ЛСО, которая не зависит от языковых показателей отношений; с другой стороны, появляется возможность описать случаи эксплицитной реализации каждого отношения (т.е. при помощи коннектора – показателя соответствующего ЛСО) в текстах на трех языках, а также другие возможные способы его реализации в переводах (например, с помощью пунктуации, морфологической формы или синтаксической конструкции).

Мы предлагаем считать классификационную схему верифицируемой, если выполняются три следующих требования:

- включение в перечень любой рубрики (любое уточнение перечня) должно быть связано с аннотированием тех текстовых фрагментов, в результате анализа которых было принято решение о включении этой рубрики в перечень (или об уточнении перечня);
- в процессе экспертного формирования перечня классификационных рубрик должно эксплицироваться соответствие между каждой рубрикой и теми текстовыми фрагментами, результатами анализа которых было обосновано ее включение в перечень;
- все аннотированные многоязычные ресурсы должны сохраняться в той форме, которая в процессе разработки или применения классификации дает возможность пользователям увидеть соответствие между каждой рубрикой и обосновывающими ее аннотированными текстовыми фрагментами.

Выполнение этих требований позволит вести в сформированной базе данных поиск обосновывающих текстовых фрагментов по любой рубрике и визуализировать соответствие между ними и этой рубрикой.

По завершении проекта планируется создать интернет-сайт с универсальной классификацией ЛСО, базой данных аннотированных текстовых ресурсов и показателей ЛСО в русском, французском и итальянском языках. В этой базе данных можно увидеть соответствие между каждой рубрикой и теми текстовыми фрагментами, которые обосновывают ее включение в перечень. После завершения проекта созданный интернет-сайт будет открыт для публичного доступа. В настоящее время доступен демонстрационный фрагмент базы данных по адресу http://a179.ipi.ac.ru/PublicLingvoProjects/main.aspx.

Основная цель настоящей статьи заключается в описании предлагаемого метода и технологии автоматизированного формирования перечня классификационных рубрик на примере первой стадии построения универсальной классификации ЛСО, и в позиционировании полученных результатов в рамках проблематики построения классификационных схем ЛСО.

КЛАССИФИКАЦИОННЫЕ СХЕМЫ КАК СПОСОБ ОРГАНИЗАЦИИ ЗНАНИЙ

Исследования по автоматизации процессов извлечения, представления и организации знаний в виде классификационных схем активно ведутся на протяжении последних нескольких десятилетий [16-27]. Однако программным манифестом для исследователей, занимающихся упорядочиванием/организацией знаний (knowledge organization), стала статья К. Ньоли, опубликованная в 2008 г. [28]. В ней подчеркивается стратегическое значение этой проблематики, обусловленное развитием информационных технологий, благодаря которым процесс междисциплинарного обмена знаниями становится все более интенсивным и приобретает глобальный характер. В статье сформулированы основополагающие принципы, которым должна отвечать классификация знаний, независимо от сферы ее применения: будь то библиотечное дело, информационная наука, философия, социология, лингвистика, web-дизайн и т.д.

Представление и организация знаний связаны с решением вопросов как онтологического (подразделение на классы, типы, части; соотнесение отдельно взятого концепта с процессом, явлением или некоторым объектом), так и эпистемологического порядка (способ познания мира человеком; выбор определенных точек зрения, их сопоставление, выбор системы координат при извлечении знаний или восприятии форм их представления). Универсальность классификации предполагает отказ от устоявшегося противопоставления онтологического и эпистемологического подходов в пользу их консолидации. Согласно К. Ньоли, оба подхода должны быть ориентированы на максимальную нейтральность и конвенциональность, обеспечивая интероперабельность разных систем организации знаний, но вместе с тем предусматривая возможность отметить в них особенности, присущие изучению какого-либо феномена в рамках заданной научной дисциплины. Кроме того, процесс организации знаний существенно зависит от временного фактора: со временем классификационные схемы могут претерпевать качественные изменения.

К. Ньоли подчеркивает, что, как и сами знания, системы их упорядочивания не могут быть одномерными. Так, при организации документов важно учитывать: 1) материальную форму, 2) онтологический статус описываемого феномена, 3) используемые узко-дисциплинарные и общетеоретические подходы, 4) выражаемую авторскую точку зрения, 5) индивидуальные предпочтения будущих пользователей, 6) изменения систем знаний.

Вместе с тем, современные системы организации знаний, как утверждает Ньоли, в большинстве своем являются одномерными (или «плоскостными») и пытаются вместить упорядочиваемые знания в одно-

или двухуровневые классификации, значительно упрощая сложную природу описываемых феноменов. Избежать этого позволяет фасетный принцип классификации, который был предложен Ш.Р. Ранганатаном [29] и явился последовательным развитием идей Сэйерса. Этот принцип обладает значительным исследовательским потенциалом и удовлетворяет поисковым запросам в ряде контекстов, связанных с проблемами комплексного описания объектов [30, 31].

Фасетный подход является сейчас одним из самых продуктивных при построении систем организации знаний [32–34]. Однако его критикуют за то, что он основывается на допущении, будто отношения между концептами существуют априори, а не устанавливаются в результате формирования и развития моделей организации знаний [35, с. 545]. Поэтому некоторые критики фасетного анализа при разработке автоматизированных систем организации знаний делают выбор в пользу формального концептуального анализа (Formal Concept Analysis – FCA/ФКА) и его логического аппарата. Они признают ключевую роль таксономий, классификаций и других онтологий как средств организации конвенциональных знаний в отдельно взятой предметной области. При этом онтология представляется ими как формальная спецификация результатов процесса концептуализации, которая понимается не как генерация знаний в рамках этой предметной области, а как ее абстрактная репрезентация в виде формальных моделей.

В своих исследованиях сторонники формального концептуального анализа (ФКА), на котором основаны технологии компьютерной категоризации, предпринимают попытку максимально автоматизировать построение таксономий, мотивируя это стремлением сократить временные издержки. В ФКА при обработке текстового материала используются методы генеративной грамматики, которые основываются на допущении о том, что глагол накладывает строгие сочетаемостные ограничения на свои аргументы. Иерархия концептов строится с учетом степени выраженности и общности этих ограничений у различных аргументов глагола [13, 14].

В то же время, использование глагольных вершин предложения в качестве основного носителя понятийного содержания концептов и показателя отношений между ними вводит значительные ограничения на обрабатываемый языковой материал. Как правило, при ФКА анализируемый отрезок не может быть больше одного предложения, несмотря на то, что имеются отдельные, хотя и не очень успешные, попытки выйти за границу предложения. Кроме того, обычно подобные исследования проводятся на моноязыковом материале, который представлен узкоотраслевыми текстами с ограниченным лексическим и синтаксическим репертуаром, а если привлекаются корпусные данные, то небольшого объема [15].

Компьютерные технологии, в отличие от экспертных, ориентированы, в основном, на создание таксономий концептов, выражаемых существительными или именными группами. В качестве исходной информации используются, в основном, тематические корпуса текстов (туризм, финансы, медицина и т.д.). Пример результата компьютерной обработки текстов по туристической тематике приведен в табл. 1, на основе которой была построена иерархическая таксономия концептов по этой тематике (рис. 1).

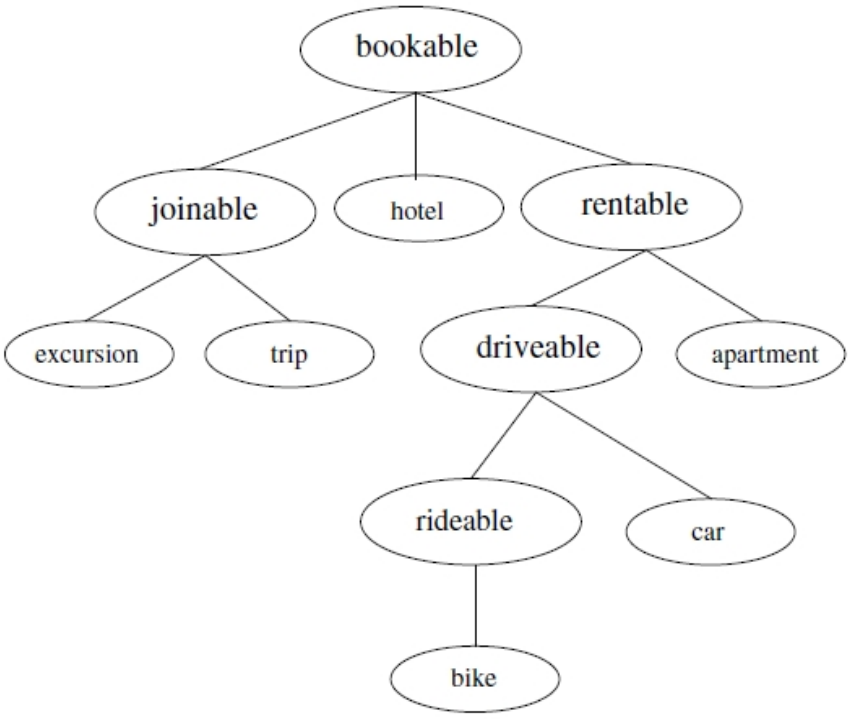

Рис. 1. Автоматически сформированная таксономия концептов и отношений между ними [14]

Таблица 1

Пример результата компьютерной обработки текстов [14]

Концепты предметной области	Свойства концептов				
	bookable	rentable	driveable	rideable	joinable
hotel	x	-	-	-	-
apartment	x	x	-	-	-
car	x	x	x	-	-
bike	x	x	x	x	-
excursion	x	-	-	-	x
trip	x	-	-	-	x

Однако для построения классификации логико-семантических отношений, которые во многих случаях устанавливаются между самостоятельными предложениями или даже между более значительными по объему фрагментами текста, эти технологии не подходят.

Современные исследования ЛСО, связанные с их классификацией, строятся на контрастивном языковом материале большого объема и привлекают широкий методологический арсенал, используя, наряду с семантическим, когнитивным и прагматическим анализом, наработки информационной науки [36]. При построении сложных классификационных схем, предполагающих многоуровневое аннотирование с учетом множественных признаков, в таких исследованиях широко применяется фасетный подход [9, 37], на котором основан и представляемый в настоящей статье проект.

ЭКСПЕРТНЫЕ ТЕХНОЛОГИИ

Во Введении нами было предложено разделить существующие технологии построения или уточнения таксономий и классификационных схем на две категории: экспертные и компьютерные, каждая из которых может быть одно- или многоязычной.

Одна из наиболее известных *экспертных одноязычных технологий* использовалась при формировании Penn Discourse Treebank (PDTB) [9]. Она была разработана для экспертного аннотирования дискурсивных отношений и их аргументов как внутри предложения, так и между предложениями. Экспертами было выполнено аннотирование текстов из корпуса «Wall Street Journal», объем которого превышает 1 млн словоупотреблений. При аннотировании эксперты различали отношения, выраженные специализированными лексическими показателями, т. е. коннекторами (explicit relations), и лексически не выраженные (implicit relations), как, например, в следующей фразе:

“But a few funds have taken other defensive steps. Some have raised their cash positions to record levels. <*Implicit = BECAUSE*> High cash positions help buffer a fund when the market falls”.

Всего экспертами было описано 18459 случаев с ЛСО, маркированными коннекторами, и 16224 случая с немаркированными ЛСО. Кроме того, поскольку многие коннекторы полисемичны, т.е. один и тот же коннектор может быть показателем разных отношений, и в некоторых случаях определить однозначно отношение невозможно, это отмечается в процессе аннотирования с помощью специальных меток (тэгов) отношений. Все метки ЛСО образуют трехуровневую иерархию: от более абстрактного смысла ЛСО к более конкретному (рис. 2).

Так, в рубрике первого уровня временных отношений (Temporal) метка второго уровня различает синхронные временные отношения (Synchronous), когда две ситуации имеют место одновременно, и асинхронные (Asynchronous), внутри которых на третьем уровне различаются предшествование (Precedence) и следование (Succession).

С помощью системы меток эксперты аннотировали, например, три разных случая использования коннектора *since* следующим образом [9]:

1. The Mountain View, Calif., company has been receiving 1,000 calls a day about the product since it was demonstrated at a computer publishing conference several weeks ago <*Temporal.asynchronous*>.

2. It was a far safer deal for lenders since NWA had a healthier cash flow and more collateral on hand <*Contingency.cause*>.

3. Domestic car sales have plunged 19% since the Big Three ended many of their programs Sept. 30 <*Temporal.asynchronous + Contingency.cause*>.

В этих примерах для аннотирования отношений экспертами были использованы метки рубрик второго уровня Temporal.asynchronous и Contingency.cause. В случае, когда эксперт не может однозначно определить выражаемое коннектором ЛСО, технология позволяет использовать одновременно две метки, как это сделано в 3-м примере (Temporal.asynchronous + Contingency.cause).

С помощью такой экспертной технологии произведено аннотирование, кроме английского, еще пяти корпусов текстов: на арабском, китайском, турецком, чешском языках, а также на языке хинди. Отметим, что корпуса текстов на перечисленных языках не являются параллельными и, следовательно, такая экспертная технология – одноязычна.

Экспертная многоязычная технология для изучения отдельных коннекторов (но не для описания системы ЛСО) рассмотрена в [11], где приведены

результаты аннотирования тех фрагментов текста на английском языке из корпуса Europarl Direct (https://www.idiap.ch/dataset/europarl-direct), которые содержат коннекторы *also, because, although* и *if*. Одновременно аннотировались переводы этих фрагментов на четыре языка (нидерландский, немецкий, французский и испанский) и отмечались случаи эксплицитного перевода анализируемых английских коннекторов в сопоставляемых языках и случаи имплицирования коннектора. Параллельные фрагменты текстов, в которых коннектор оригинального текста не был переведен, исключались из рассмотрения. Соотношение случаев имплицирования и эксплицирования коннекторов в параллельных текстах на анализируемых языках представлено в табл. 2.

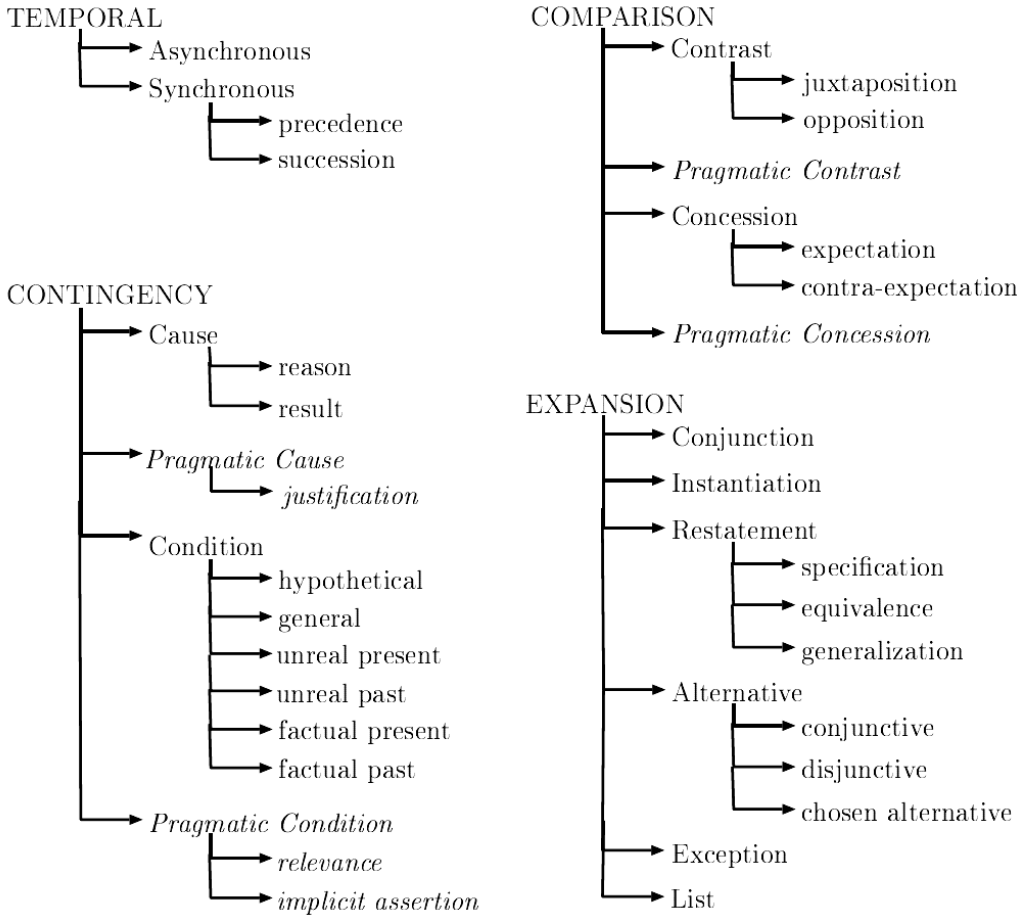

Рис. 2. Классификационная схема логико-семантических отношений [9, с. 5]

Таблица 2

Соотношение случаев эксплицирования и имплицирования коннектора [11]

Языки перевода:	**Нидерландский**	**Немецкий**	**Французский**	**Испанский**
Also	19/194	15/181	14/190	7/195
Because	27/383	8/391	19/389	4/393
Although	7/248	2/248	5/256	1/261
If	0/226	0/201	0/222	0/241

Примечание:

перед символом «/» указано количество переводов, где один из четырех коннекторов, указанных в первом столбце, не имеет лексического выражения (для переводов *also* на нидерландский язык – 19);

после символа «/» указано количество переводов, в которых четыре английских коннектора переведены соответствующим коннектором в сопоставляемых языках или другими средствами выражения ЛСО, кроме парафраз и синтаксических конструкций (для переводов *also* на нидерландский язык – 194).

Экспертная многоязычная технология уточнения применяемой при формировании PDTB системы логико-семантических отношений [9] рассмотрена в работе [36], в которой приведены результаты аннотирования текстов и их переводов на пяти языках (английский, французский, немецкий, нидерландский и итальянский). В качестве компарандума (pivot language) был выбран английский язык (но он при этом не всегда являлся языком оригинала), и аннотировались все случаи использования каждого из 36 английских коннекторов[1], а также их переводные эквиваленты в языках-компаратумах, кроме парафраз. По итогам этого эксперимента, а также в связи с теоретическими проблемами определения ЛСО в используемой PDTB классификации, была предложена модифицированная классификационная схема для ЛСО [36, с. 15].

Помимо уже упомянутых, существует еще несколько классификационных схем ЛСО [2, 38, 39]. Почему возникла необходимость в разработке еще одной классификации?

Прежде всего, существующие классификации имеют два уровня и на каждом может быть установлено ЛСО: в классификации, предложенной в [39], это уровень пропозиционального содержания и уровень структуры текста, т.е. все те отношения, которые «комментируют» выбор формы высказывания говорящим или место вводимого им фрагмента текста в структуре текста. В классификации, используемой в PDTB, тоже два уровня, но их выбор отличается от приведенного выше: это уровень пропозиционального содержания и уровень речевого акта (ср. оппозицию pragmatic vs. non-pragmatic на рис. 2: в группе CONTINGENCY различаются Cause и Pragmatic Cause, в группе COMPARISON Contrast и Pragmatic Contrast, Concession и Pragmatic Concession). Предлагаемая нами классификация (подробнее [40, 41]) последовательно различает три уровня: пропозициональное содержание, речевой акт и структура текста, допуская, в отличие от [39], что одно и то же отношение может быть теоретически установлено на всех трех уровнях [42]. Так, отношение коррекции может быть установлено на уровне пропозиционального содержания связываемых этим отношением фрагментов текста (в высказывании *Он приехал не на машине, а на велосипеде* положение вещей *Он приехал на машине* замещается другим *Он приехал на велосипеде*); оно может быть установлено на уровне речевого акта (*Он очень умный... то есть... мне так кажется*, где фрагмент текста, вводимый *то есть*, корректирует категоричность речевого акта утверждения, совершенного при помощи первого фрагмента текста); наконец, отношение коррекции может быть метатекстовым, т.е. касаться выбора языковых средств для описания одной и той же ситуации (например, *Она не толстая, а полноватая*).

Кроме того, в предлагаемой нами классификации, также исходящей на первом уровне из наиболее абстрактного смысла, общего для некоторой группы отношений, в основу положены четыре группы семантических операций: импликация, сравнение, расположение на хронологической оси и соотнесение части и целого или элемента и множества. Последняя операция не учитывается ни в одной из предложенных классификаций, но позволяет увидеть тот общий семантический механизм, который связывает отношения генерализации, спецификации, исключения и включения, объединенные на этом основании в группу мереологических отношений [41]. В существующих классификациях эти отношения находятся, как правило, в разных группах (например, рис. 2, где отношения генерализации и спецификации отнесены к рубрике второго уровня «restatement – переформулирование», а отношение исключения позиционируется как рубрика второго уровня, не связанная с отношениями генерализации и спецификации).

Наконец, сочетание двух критериев – уровень, на котором установлено ЛСО, и семантическая операция, лежащая в его основе, – позволяет увидеть не только то общее, что характеризует ту или иную группу отношений, но и то, что отношения, основанные на разных семантических операциях, но установленные на одном семантическом уровне, могут выполнять одну и ту же функцию в тексте. Например, отношения спецификации, переформулирования и альтернативы, установленные на метаязыковом уровне, т.е. на уровне структуры текста, в равной степени могут быть использованы с пояснительной функцией.

Таким образом, предлагаемая классификация логико-семантических отношений представляет собой не просто локальное уточнение существующих схем, а является принципиально новой. Ее построение потребует аннотирования существенно более представительных параллельных корпусов (как по объему, так и по жанровому многообразию), чем те, что представлены в [36], где каждый из пяти параллельных газетных текстов содержал приблизительно 8500 словоупотреблений. Предметом аннотирования были 36 коннекторов английского языка и их переводные эквиваленты. В тексте на английском языке они встретились 203 раза. При этом ЛСО второго уровня «restatement – переформулирование», включающее отношения генерализации и спецификации, аннотировано 4 раза, отношение исключения не встретилось ни разу.

Кроме того, количество показателей ЛСО, которые аннотируются в нашем проекте, превышает более чем в шесть раз число таких показателей в PDTB – одном из самых больших аннотированных одноязычных корпусов: более 100 и более 600, соответственно. Общее число словоупотреблений в используемых нами текстах Национального корпуса русского языка (более 2,5 млн) в 2,5 раза превышает объем текстов PDTB (1 млн). Существенно бо́льший объем текста на русском языке обеспечивает представительное число аннотированных показателей ЛСО для каждой классификационной рубрики как в основном тексте, так и в его переводах.

[1] Аннотировались следующие 36 коннекторов (в скобках указаны их частотности в английских текстах корпуса): *after* (1), *even if* (4), *in short* (1), *then* (3), *although* (6), *for example* (3), *in spite of* (1), *therefore* (3), *and* (50), *for instance* (1), *indeed* (1), *though* (5), *as* (3), *given (that)* (2), *meanwhile* (1), *thus* (2), *as well as* (1), *however* (7), *now* (2), *well* (1), *because* (5), *if* (11), *or* (5), *when* (7), *before* (4), *in fact* (1), *since* (1), *whether* (2), *but* (41), *in order to* (1), *so* (2), *while* (9), *despite* (6), *in other words* (1), *that is why* (1) и *yet* (8).

Итак, в нашем проекте поставлена задача построения новой универсальной классификации логико-семантических отношений между частями предложений, предложениями и фрагментами параллельных текстов. С этой целью мы предлагаем принципиально новые метод и экспертную многоязычную технологию построения классификационных схем ЛСО, обеспечивающие возможность их формирования практически с «чистого листа».

МЕТОД И ТЕХНОЛОГИЯ

Предлагаемый метод формирования перечня классификационных рубрик заключается в аннотировании экспертами языковых объектов или явлений как в моноязычных, так и в параллельных выровненных текстах на двух или более языках, на основе одних и тех же принципов использования инструментальных фасетных классификаций (ИФК), формируемых в процессе аннотировании. Для текстов каждого языка применяется своя ИФК, причем, в начале аннотирования используемые ИФК могут быть пустыми или заполненными частично. После завершения процесса аннотирования на основе рубрик всех сформированных инструментальных фасетных классификаций экспертами создается целевая классификация исследуемых языковых объектов или явлений.

Начнем описание предлагаемого метода формирования перечня классификационных рубрик с примера аннотирования параллельных выровненных текстов[2] на русском и французском языках с применением двух ИФК. Одна классификация применяется в процессе аннотировании русских текстов (ее обозначим ИФК1), а вторая – их переводов на французский язык (ИФК2). Затем опишем метод формирования в более общем виде.

Этот метод, ориентированный на компьютерную реализацию, может применяться как при сплошном аннотировании исследуемых информационных объектов (т.е. от начала до конца текста [9]), так и при выборочном их аннотировании [11]. Во втором случае эксперты сначала формируют массив пар текстовых фрагментов из одного или нескольких предложений, задавая некоторый критерий их отбора. На рис. 3 показаны первые четыре пары из массива, сформированного по критерию наличия многоместного сочетания *не только …, но* в русских текстах. Первые две пары содержат сочетание *не только …, но и*, четвертая – *не только …, но*, а третья – более сложное сочетание *не только..., но даже..., а иногда и*.

Если используются параллельные тексты, то при любом принципе аннотирования (сплошном или выборочном) объектом экспертного семантического анализа являются одновременно левый и правый текстовые фрагменты каждой пары. Как было отмечено, в общем случае предметом аннотирования может быть любой информационный объект моноязычных или параллельных текстов. Здесь предлагаемый метод формирования перечня классификационных рубрик описывается на примере выборочного аннотирования реализаций ЛСО в параллельных текстах. Результат применения этого метода в четвертой паре на рис. 3 имеет вид табл. 3. В единичных случаях подобные таблицы могли бы быть сформированы и без применения средств информатики. Но при обработке текстовых массивов объемом в миллионы словоупотреблений формирование тысяч и десятков тысяч таких таблиц предполагает компьютерную реализацию предлагаемого метода.

Содержание столбцов 2 и 4 табл. 3 отличается только первой рубрикой при совпадении остальных, но в общем случае рубрики могут быть разными. Описание содержания, процессов наполнения и применения обеих ИФК будет приведено далее при описании метода формирования перечня рубрик целевой классификации.

Предлагаемый метод позволяет строить и применять одновременно N ИФК, где N – число языков, тексты на которых аннотируются для построения целевой классификации. В каждой ИФК эксперты могут выделить главный фасет и несколько дополнительных. В этом случае рубрики главного фасета выделяются полужирным шрифтом (табл. 3). Общее число фасетов любой ИФК может изменяться экспертами в процессе аннотирования исследуемых информационных объектов. Рубрики каждого фасета ИФК могут быть представлены в виде списка или многоуровневой иерархии. В последнем случае количество уровней и число рубрик ИФК на каждом уровне также могут изменяться экспертами в процессе аннотирования, т.е. в общем случае все ИФК являются темпоральными и по структуре, и по наполнению их рубриками.

В примере с аннотированием языковых реализаций ЛСО эксперты использовали семифасетную ИФК1 для русских текстов и восьмифасетную ИФК2 – для их переводов на французский язык, выделив в них главные двухуровневые иерархические фасеты для классификации коннекторов в русском и французском языках. Первый (верхний) уровень главного фасета ИФК1 включает 4 рубрики, а ИФК2 – 7. Названия рубрик первого уровня, а также число рубрик второго уровня для каждой из них приведены в табл. 4 и 5, соответственно. Число рубрик на втором уровне указано по состоянию на август 2016 г.

По главному фасету ИФК1 двухместное сочетание *не только …, но и* в первых двух парах на рис. 3 отнесено экспертами к рубрике второго уровня **не только||но и** двухместное сочетание *не только …, но* в четвертой паре отнесено к рубрике **не только||но**, а многоместное сочетание *не только..., но даже..., а иногда и* в третьей паре – к рубрике **не только||но даже||а иногда и**. По главному фасету ИФК2 двухместный коннектор *non seulement …, mais* в четвертой паре на рис. 3 отнесен к рубрике второго уровня **non seulement||mais** (см. табл. 3). Если необходимые рубрики главных и/или дополнительных фасетов отсутствуют в ИФК, то эксперты могут создавать новые рубрики и редактировать их в процессе аннотирования, что является отличительной чертой предлагаемого метода. На середину августа 2016 г. эксперты сформировали около 6500 аннотаций. Из них в 25 аннотациях использовалась рубрика **не только||но,** в 53 – **не только||но и** и в 38 – **non**

[2] Для примеров используются фрагменты текстов параллельного французско-русского подкорпуса Национального корпуса русского языка (НКРЯ). Общий объем этого подкорпуса составляет 2,5 млн словоупотреблений.

seulement||mais. При этом в 11 аннотациях двухместное сочетание *не только …, но* было переведено как *non seulement …, mais*.

Кроме рубрик главных фасетов, эксперты при аннотировании применяют рубрики дополнительных фасетов. В 2016 г. эксперты использовали шесть дополнительных фасетов в ИФК1 и семь – в ИФК2. Названия дополнительных фасетов ИФК1 и количество их рубрик приведены в табл. 6. Они совпадают с первыми шестью дополнительными фасетами ИФК2 (рубрики седьмого фасета ИФК2 используются для классификации ошибок машинного перевода). Отметим, что названия всех фасетов также выбираются экспертами и могут редактироваться ими в процессе аннотирования. Иначе, предлагаемый метод формирования перечня классификационных рубрик не зависит от выбранных экспертами названий фасетов и их рубрик.

Оригинальный текст	Перевод
Захар не старался изменить **не только** данного ему Богом образа, **но и** своего костюма, в котором ходил в деревне.	Zakhar n'avait rien fait pour changer l'apparence que Dieu lui avait donnée ni le costume qu'il avait porté à la campagne.
Он его представлял себе чем-то вроде второго отца, который только и дышит тем, как бы за дело и не за дело, сплошь да рядом, награждать своих подчиненных и заботиться **не только** о их нуждах, **но и** об удовольствиях.	Il se l'imaginait comme une sorte de second père qui ne pensait qu'à distribuer des primes à ses employés, qu'ils le méritent ou non, à tort et à travers, et qu'à veiller non seulement à leurs besoins mais aussi à leur bien-être.
Это происходило, как заметил Обломов впоследствии, оттого, что есть такие начальники, которые в испуганном до одурения лице подчиненного, выскочившего к ним навстречу, видят **не только** почтение к себе, **но даже** ревность, **а иногда и** способности к службе.	Comme Oblomov le remarqua plus tard, la cause en était que certains supérieurs voyaient dans la mine effrayée d'un employé qui s'empressait à leur rencontre, non seulement une preuve de respect pour eux, mais aussi un signe de zèle et même d'aptitude au service.
Со времени смерти стариков хозяйственные дела в деревне **не только** не улучшились, **но**, как видно из письма старосты, становились хуже.	Depuis la mort des parents, les affaires du domaine non seulement ne s'amélioraient pas, mais, à en croire la lettre du régisseur, empiraient.

Рис. 3. Четыре пары выровненных предложений с многоместным сочетанием *не только …, но*

Таблица 3

Результат одновременного аннотирования реализации ЛСО, выраженного *не только…, но*, и ЛСО в переводе с его функциональным эквивалентом *non seulement… mais*

Со времени смерти стариков хозяйственные дела в деревне **не только** не улучшились, **но**, как видно из письма старосты, становились хуже.	**не только||но** < CNT > < CNT p CNT q > < Дистант > < неединственности >	Depuis la mort des parents, les affaires du domaine **non seulement** ne s'amélioraient pas, **mais**, à en croire la lettre du régisseur, empiraient.	**non seulement||mais** < CNT p CNT q > < CNT > < Дистант > < неединственности >

Примечание: Первый столбец табл. 3 содержит *не только…, но* в функции коннектора в контексте, необходимом для его анализа. Второй столбец содержит пять рубрик ИФК1, присвоенных экспертами этому коннектору. Третий столбец содержит *non seulement… mais*, его функционально-эквивалентный фрагмент (ФЭФ), а также перевод контекста на французский язык, необходимого для анализа французского коннектора. Четвертый столбец содержит пять рубрик ИФК2, присвоенных экспертами этому функционально-эквивалентному фрагменту.

Таблица 4

Названия четырех рубрик первого уровня главного фасета ИФК1

№	Рубрика первого уровня для классификации коннекторов русского языка	Число рубрик второго уровня
1	Однокомпонентные	56
2	Многокомпонентные	319
3	Двухместные	189
4	Многоместные	22

Таблица 5

Названия семи рубрик первого уровня главного фасета ИФК2

№	Рубрика первого уровня для классификации ФЭФ французского языка	Число рубрик второго уровня
1	Однокомпонентные	41
2	Многокомпонентные	382
3	Двухместные	145
4	Многоместные	17
5	Отрицание	110
6	Error	65
7	Прочие ФЭФ	205

Таблица 6

Названия дополнительных фасетов ИФК1 и количество их рубрик по состоянию на август 2016 г.

№	Фасет	Количество рубрик
1	Отношения	23
2	Структура	8
3	Позиция	3
4	Порядок	9
5	Статус	5
6	Расположение	2

Для построения целевой классификации ЛСО ключевым является фасет «Отношения», который представляет собой регулярно пополняемый список классификационных рубрик. По состоянию на 31 декабря 2015 г. в нем было 10 рубрик, включая специальную рубрику «Отношение подлежит определению – to be defined (**TBD**)», куда отнесены те аннотации, где ЛСО не могут быть описаны девятью другими рубриками, или те аннотации, где ЛСО требует уточнения. Всего была сформирована 991 аннотация. Для 474 из них была проставлена рубрика ЛСО, включая 381 позицию **TBD** [43].

По состоянию на 30 июня 2016 г. общее число сформированных аннотаций увеличилось до 4202, из них для 2618 проставлены рубрики ЛСО, количество которых увеличилось до 23, включая **TBD** (табл. 7).

При компьютерной реализации метода формирования перечня классификационных рубрик количество аннотаций для каждой ЛСО, кроме нуля, является гиперссылкой, по которой эксперт, а затем и пользователь, может увидеть соответствующие аннотированные реализации ЛСО. Например, в третьем столбце табл. 7 число аннотаций с ЛСО «Отношение аналогии» равно 2. Обратившись по гиперссылке «2», эксперт увидит на экране две аннотации (табл. 8). Отметим, что нули в табл. 7 говорят о том, что соответствующие ЛСО ни разу не были проставлены в сформированных экспертами аннотациях, т.е. на момент начала аннотирования параллельных текстов фасет «Отношения» был заполнен частично.

В правой части обеих аннотаций (см. табл. 8) проставлены рубрики дополнительных фасетов из табл. 6, которые были описаны в [43]. Поэтому ограничимся кратким описанием содержания этих шести фасетов, используя в качестве примера вторую аннотацию из табл. 8:

- рубрика <сравнительные> относится к фасету «Отношения» (см. строку 1 в табл. 6) и говорит о том, что коннектор *comme* выражает в переводе сравнительное отношение (коннектор *как… так… так… и* выражает в оригинале отношение аналогии);
- рубрика <без предикации> относится к фасету «Структура» (см. строку 2 в табл. 6) и говорит о том, что коннектор *comme* маркирует часть предложения без предикации;
- рубрика <начальная> относится к фасету «Позиция» (см. строку 3 в табл. 6) и говорит о том, что коннектор *comme* стоит в начальной позиции в маркируемом им фрагменте текста;

Таблица 7

Названия 23 рубрик фасета «Отношения» ИФК1 и количество аннотаций по каждой рубрике

№	Название ЛСО	Количество аннотаций	
		на 31.03.2016 г.	на 30.06.2016 г.
1	Временные отношения	0	287
2	Отношение альтернативы	0	1
3	Отношение аналогии	0	2
4	Отношение генерализации	0	190
5	Отношение исключения	0	16
6	Отношение неединственности	16	133
7	Отношение подлежит определению	807	1311
8	Отношение противопоставления	1	131
9	Отношение следствия	1	1
10	Отношение спецификации	0	32
11	Отношение замещения	35	38
12	Отношение коррекции	0	20
13	Отрицание тождества	2	2
14	Присоединительные отношения	0	5
15	Причинно-следственные отношения	1	45
16	Причинные отношения	0	3
17	Противительные отношения	22	87
18	Соединительные отношения	4	4
19	Сопоставительные отношения	0	4
20	Сравнительные отношения	0	4
21	Условные отношения	0	264
22	Уступительные отношения	10	38
23	Целевые отношения	0	0

Таблица 8

Две аннотации с ЛСО «Отношение аналогии» для двух французских переводов одного и того же предложения на русском языке

Как что делалось при дедах и отцах, **так** делалось при отце Ильи Ильича, **так**, *может быть*, делается еще **и** теперь в Обломовке.	**как||так||так||и** < CNT > < Дистант > < CNT p CNT q [= CNT r, CNT s ...] > < аналогия >	A l époque du père d Ilia Ilitch tout allait **comme** du temps des aïeuls Probablement qu aujourd'hui rien n a encore changé à Oblomovka	**comme** < начальная > < без предикации > < p CNT q > < CNT > < сравнительные >
Как что делалось при дедах и отцах, **так** делалось при отце Ильи Ильича, **так**, *может быть*, делается еще **и** теперь в Обломовке.	**как||так||так||и** < CNT > < Дистант > < CNT p CNT q [= CNT r, CNT s ...] > < аналогия >	*Bref*, tout se faisait, du temps d'Ilia Ilitch, **comme** du temps de son père et de son arrière-grand-père...	**comme** < начальная > < без предикации > < p CNT q > < CNT > < SubCNT > < сравнительные >

- рубрики < CNT p CNT q [= CNT r, CNT s ...] > и < p CNT q > относятся к фасету «Порядок» (см. строку 4 в табл. 6) и говорят о порядке следования вводимых коннектором фрагментов текста;
- рубрики <CNT> и <SubCNT> отноятся к фасету «Статус» (см. строку 5 в табл. 6) и говорят о том, что аннотация создана для всего коннектора *как ... так ... так и* в целом (<CNT>)[3], и что данный коннектор является встроенным (< SubCNT >), т.е. находится в сфере действия другого коннектора;
- рубрика <Дистант> относится к фасету «Расположение» (см. строку 6 в табл. 6) и говорит о том, что компоненты коннектора *как ... так ... так ... и* разделены текстом.

Рассмотренный пример аннотирования реализаций ЛСО в параллельных текстах является наглядной иллюстрацией пополнения и использования фасет ИФК. Предлагаемый подход к их пополнению позволяет применять предлагаемый метод формирования классификационных схем практически с «чистого листа» в процессе аннотирования исследуемого вида информационных объектов в моноязыковых и параллельных текстах.

В одних и тех же текстах с помощью этого метода могут одновременно аннотироваться несколько видов информационных объектов разными коллективами экспертов, решающих разные лингвистические задачи [4–8]. При этом эксперты каждого коллектива могут визуализировать аннотации только своего вида информационных объектов. Эта возможность одновременного, но раздельного аннотирования нескольких видов объектов определяет первое положение метода и, соответственно, требование к его компьютерной реализации.

Положение 1. Предлагаемый метод не предполагает, что экспертная разметка, аннотации информационных объектов и их контекстов хранятся в обрабатываемых текстах, но он обеспечивает связанность разметки с аннотируемыми информационными объектами и их контекстами с помощью системы ссылок.

Необходимость обеспечения верифицируемости определяет 2-е положение метода и требование к его компьютерной реализации.

Положение 2. Метод связывает сформированные экспертами аннотации, с одной стороны, с информационными объектами исследуемого вида и их контекстами системой ссылок, а с другой стороны – с ИФК, формируемыми в процессе аннотирования для каждого вида исследуемых объектов.

Обобщая рассмотренный пример аннотирования реализаций ЛСО, сформулируем еще ряд положений.

Положение 3. Метод предполагает использование классификаций, которые формируются для исследуемых объектов каждого вида и являются, в общем случае, фасетными и темпоральными, а также дает возможность их редактирования в процессе экспертного аннотирования. Рубрики одной из фасет предназначены для формирования целевой классификации исследуемых объектов.

Положение 4. При контрастивном анализе информационных объектов в параллельных текстах на двух и более языках метод обеспечивает экспертам возможность формирования аннотаций для каждого из анализируемых языков за счет использования системы ссылок к аннотированным объектам исследуемого вида, их функционально-эквивалентных фрагментов, контекстов на исследуемых языках, а также за счет ссылок к соответствующим ИФК, количество которых равно числу языков. В процессе аннотирования формируется и система обратных множественных ссылок от рубрик ИФК каждого языка к аннотациям для других исследуемых языков, объектам исследуемого вида, функционально-эквивалентным фрагментам и их контекстам на том языке, для которого формируется эта ИФК.

Положение 5. В процессе контрастивного анализа метод обеспечивает возможность аннотирования экспертами информационных объектов при наличии нескольких переводов текста, в котором они встречаются, за счет предварительного выравнивания текста и его переводов.

Положение 6. Метод при его компьютерной реализации может применяться как для сплошного аннотирования информационных объектов в каждом тексте, так и для выборочного их аннотирования. Во втором случае эксперты сначала формируют массив текстовых фрагментов из одного или нескольких предложений с информационным объектом исследуемого вида, задавая критерий их отбора.

Положение 7. После завершения сплошного или выборочного аннотирования информационного объекта в моноязычных текстах, рубрики одной из фасет представляют собой перечень классификационных рубрик, предназначенных для построения экспертами целевой классификации исследуемых объектов. Для параллельных текстов перечень классификационных рубрик строится как объединение рубрик соответствующих фасет всех ИФК, сформированных при аннотировании информационного объекта исследуемого вида.

Перечисленные семь положений метода формирования перечня классификационных рубрик легли в основу разработки экспертной многоязычной технологии.

Однако необходимо сделать одно уточнение. Из Положения 7 следует, что для формирования перечня используются рубрики только одного фасета каждой ИФК. В примере с аннотированием реализаций ЛСО эксперты использовали семифасетную ИФК1 для русских текстов и восьмифасетную ИФК2 – для французских. Для формирования перечня ЛСО достаточно объединить рубрики фасет «Отношения» ИФК1 и ИФК2, учитывая, что они во многом будут совпадать.

Однако остальные фасеты дают возможность сформировать в целевой классификации для каждого ЛСО (а не только для его реализаций) многоаспектную аннотацию. Например, для ЛСО «Причинно-следственные отношения» можно указать не только общее число найденных для него языковых реализаций на некоторый момент времени аннотирования или после его завершения, но и их распределение по рубрикам реализаций ЛСО в текстах каждого языка.

[3] Используемый метод позволяет аннотировать составляющие коннектора и их переводы [43].

Приведем в табл. 9 данные о распределении 45 реализаций ЛСО «Причинно-следственные отношения» на 30.06.2016 г. (см. строку 15 в табл. 7) по рубрикам главного фасета ИФК1.

Таблица 9

Распределение 45 реализаций ЛСО «Причинно-следственные отношения»

Рубрики главного фасета ИФК1	Число реализаций ЛСО
так как || то	29
как || то	7
коли || то значит	3
когда || то	2
так как || то потому	1
так как || так	1
когда || когда || то	1
когда || тогда	1
Итого реализаций ЛСО	45

Примечание: Символ || обозначает наличие текста между компонентами коннектора

Таким образом, семантическая разметка показателя ЛСО и использование рубрик фасет ИФК1 позволяет экспертам заполнить левую составляющую аннотации, анализируя оригинальные предложения, а семантическая разметка переводного эквивалента показателя ЛСО и использование рубрик фасет ИФК2 – правую составляющую аннотации, анализируя их переводы, что дает возможность сформировать в целевой классификации многоаспектные аннотации для всех рубрик показателей ЛСО.

На основе описанного метода была разработана экспериментальная экспертная многоязычная технология, обеспечивающая формирование перечня верифицируемых рубрик для классификации ЛСО в лингвистике с помощью надкорпусной базы данных (НБД) аннотаций [43], ИФК и параллельных текстов Национального корпуса русского языка (см. сноску 2). Аннотация формируется в результате экспертной разметки показателя ЛСО, каждый из которых на рис. 3 выделен полужирным шрифтом, и соответствующего контекста, разметка которого на рис. 3 не показана.

Разработанная технология формирования перечня классификационных рубрик выполняется итерационно. Результатом итерации является новая аннотация или ее уточнение, а после завершения аннотирования формируется перечень классификационных рубрик. При каждой итерации последовательно выполняются следующие операции:

- из параллельных текстов НКРЯ отбирается последовательно или выборочно согласно заданному критерию, пара выровненных предложений, содержащая исследуемый информационный объект в левой составляющей этой пары (см. рис. 3, четыре пары которого содержат в своих левых частях показатели ЛСО)[4];
- в отобранной паре эксперты сначала отмечают левую и правую границы исследуемого информационного объекта в целом и его компонентов, а затем заполняют левую составляющую аннотации, анализируя оригинальные предложения и проставляя в ней соответствующие рубрики фасет ИФК1[5] (см. левые составляющие аннотаций в табл. 3 и 8, т. е. первые два столбца этих таблиц);
- в отобранной паре эксперты отмечают левую и правую границы переводного эквивалента исследуемого информационного объекта в целом и его компонентов, а затем заполняют правую составляющую аннотации, анализируя переводные предложения и проставляя в ней соответствующие рубрики фасет ИФК2 (см. правые составляющие аннотаций в табл. 3 и 8, т. е. последние два столбца этих таблиц)[6];
- если необходимые рубрики отсутствуют в текущих версиях ИФК1 и/или ИФК2, то в аннотации ставятся специальные пометки, которые свидетельствуют о неполноте текущих версий и о необходимости их уточнения;
- аннотация отобранной пары и ее ссылки на ИФК и параллельные тексты записываются в надкорпусной базе данных.

Для пополнения и уточнения текущих версий ИФК1 и ИФК2 служат отдельные итерации разработанной технологии, на которых происходит их редактирование для пополнения новыми рубриками, а также замены специальных пометок этими рубриками. Формирование перечня классификационных рубрик осуществляется после завершения аннотирования. Этот перечень наследует связи ИФК1 и ИФК2 с параллельными текстами и сформированными аннотациями, что условно обозначено стрелками «1» и «2» на рис. 4. Эти связи и обеспечивают верифицируемость классификационных рубрик.

Рис. 4 условно разделен на две части: его верхняя часть включает те технологические операции, которые выполняются экспертами в процессе аннотирования и формирования перечня, а нижняя – содержит те информационные ресурсы, которые обеспечивают работу экспертов.

[4] Каждая пара может включать несколько предложений оригинального текста и их переводы, так как ЛСО может связывать отдельные предложения. Эта технология применима и в тех случая, когда исследуемый информационный объект находится в правой составляющей пары, а предметом исследования являются стимулы оригинального текста, влияющие на появление исследуемого объекта в переводах.

[5] На рис. 3 исследуемые информационные объекты обозначены полужирным шрифтом.

[6] Если переводной эквивалент исследуемого информационного объекта отсутствует, то в правой составляющей аннотации ставится специальная метка *zero*.

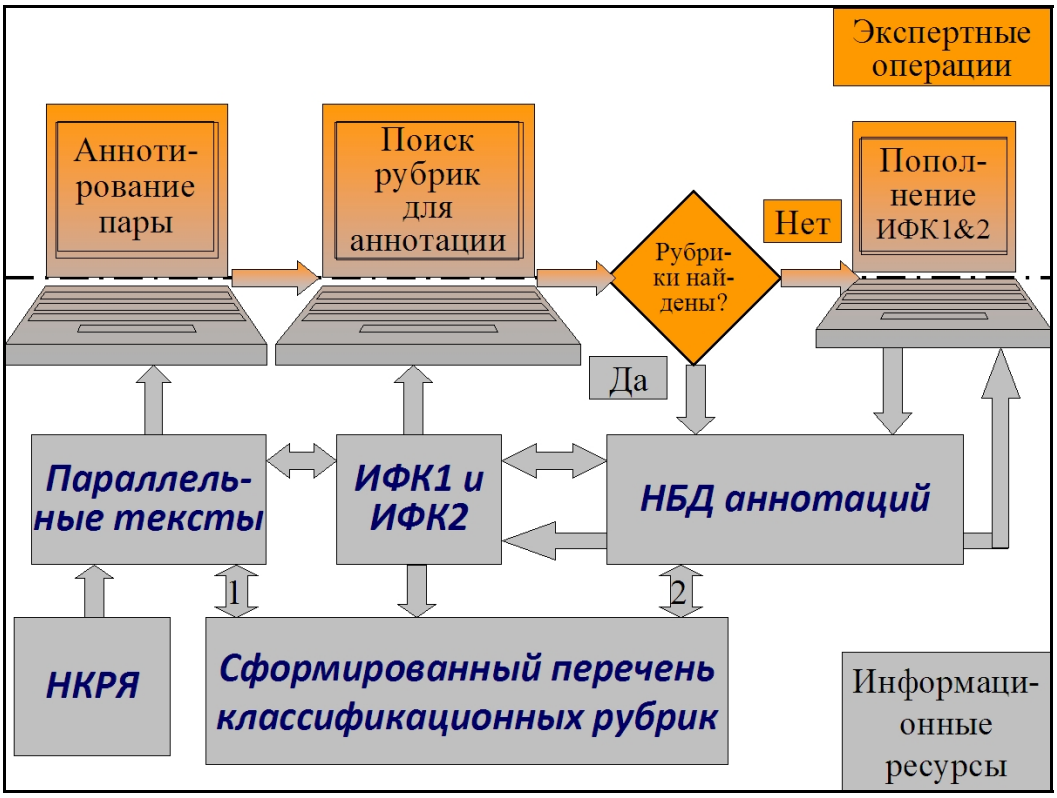

Рис. 4. Создание аннотации, пополнение инструментальных фасетных классификаций и формирование перечня классификационных рубрик

Ключевая операция разработанной технология обозначена вопросом «Рубрики найдены?». Если ответ «Да», то на соответствующей итерации в текущих версиях ИФК1 и ИФК2 уже есть рубрики, необходимые экспертам для завершения формирования аннотации информационного объекта в отобранной паре, если ответ «Нет», то необходимых рубрик нет. Поэтому в аннотации ставятся специальные пометки, которые свидетельствуют о неполноте текущих версий ИФК1 и ИФК2, а соответствующая аннотация остается сформированной только частично. На итерациях редактирования ИФК1 и ИФК2 пополняются с целью завершения формирования частично заполненных аннотаций.

ЗАКЛЮЧЕНИЕ

Предлагаемые метод и технология автоматизированного формирования перечня рубрик позволяют создавать верифицируемые классификационные схемы в широком спектре информационных объектов. В рамках рассмотренного проекта реализуемость технологии формирования перечня была проверена экспериментально в процессе аннотирования показателей логико-семантических отношений.

Проведенный эксперимент показал, что в процессе аннотирования могут формироваться перечни рубрик для создания классификаторов практически с «чистого листа». Каждая рубрика перечня логически является гиперссылкой на список тех аннотаций, в которых эта рубрика проставлена экспертами. При этом каждая аннотация имеет ссылку на контекст параллельных текстов, на основе анализа которого она была получена, что обеспечивает верифицируемость каждой из рубрик с ненулевым числом ссылок на аннотации и, соответственно, на тексты.

В существующих методах аннотирования лингвистических объектов или явлений до начала обработки параллельных текстов выбирается тот или иной вариант перечня рубрик или некоторая система классификации. После завершения работ по аннотированию выбранный перечень может модифицироваться, но в процессе аннотирования он остается неизменным [36]. Возможность формирования перечня с «чистого листа» непосредственно в процессе аннотирования является существенным элементом новизны предлагаемого метода и разрабатываемых на его основе технологий.

Таким образом, рассмотренный метод применим только на первой стадии построения классификационной схемы логико-семантических отношений: автоматизированном формировании перечня классификационных рубрик. Что касается второй стадии (создание классификационной схемы на основе сформированного перечня), то это самостоятельная задача.

СПИСОК ЛИТЕРАТУРЫ

1. Prasad R., Bunt H. Semantic relations in discourse: the current state of ISO 24617-8 // Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11). – Tilburg: Tilburg University, 2015. – P. 80–92.

2. Bunt H., Prasad R. ISO-DR-Core (ISO 24617-8): Core concepts for the annotation of discourse

relations // Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12), 2016. – P. 45-54. – URL: http://www.lrec-conf.org/ proceedings/lrec2016/LREC2016_Proceedings.zip (состояние страницы на 14.09.2016).

3. Marcus M.P., Santorini B., Marcinkiewicz M.A. Building a Large Annotated Corpus of English: The Penn Treebank // Computational Linguistics. – 1993. – Vol. 19, № 2. – P. 313–330.

4. Loiseau S., Sitchinava D.V., Zalizniak Anna A., Zatsman I.M. Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения. – 2013. – Т. 7, № 2. – С. 100–109.

5. Kruzhkov M.G., Buntman N.V., Loshchilova E.Ju., Sitchinava D.V., Zalizniak Anna A., Zatsman I.M. A database of Russian verbal forms and their French translation equivalents // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». – Вып. 13 (20). – М.: РГГУ, 2014. – С. 284–296.

6. Бунтман Н.В., Зализняк Анна А., Зацман И.М., Кружков М.Г., Лощилова Е.Ю., Сичинава Д.В. Информационные технологии корпусных исследований: принципы построения кросс-лингвистических баз данных // Информатика и ее применения. – 2014. – Т. 8, № 2. – С. 98–110.

7. Zatsman I., Buntman N. Outlining Goals for Discovering New Knowledge and Computerised Tracing of Emerging Meanings Discovery // Proceedings of the 16th European Conference on Knowledge Management. – Reading: Academic Publishing International Limited, 2015. – P. 851–860.

8. Зализняк Анна А., Зацман И.М., Инькова О.Ю., Кружков М.Г. Надкорпусные базы данных как лингвистический ресурс // Корпусная лингвистика – 2015. Труды 7-й Международной конференции. – СПб.: СПбГУ, 2015. – С. 211–218.

9. Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B. The Penn Discourse TreeBank 2.0 // Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC) / eds. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, D. Tapias. – Paris: European Language Resources Association (ELRA), 2008. – P. 2961–2968.

10. Prasad R., Webber B., Joshi A. Reflections on the Penn Discourse Treebank, Comparable Corpora, and Complementary Annotation // Computational Linguistics. – 2014. – Vol. 40, № 4. – P. 921–950.

11. Hoek J., Zufferey S. Factors Influencing the Implicitation of Discourse Relations across Languages // Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11). – Tilburg: Tilburg University, 2015. – P. 80–92.

12. Giunchiglia F., Maltese V., Madalli D., Baldry A., Wallner C., Lewis P., Denecke K., Skoutas D., Marenzi I. Foundations for the representation of diversity, evolution, opinion and bias // Technical Report DISI-09-063. – Trento: University of Trento, 2009. – URL: http://disi.unitn.it/~knowdive/ documents/foundations.pdf (состояние страницы на 14.09.2016).

13. Cimiano P., Staab S., Tane J. Automatic acquisition of taxonomies from text: FCA meets NLP // Proceedings of the PKDD/ECML'03 International Workshop on Adaptive Text Extraction and Mining (ATEM). – 2003. – P. 10–17. – URL: http://staffwww.dcs.shef.ac.uk/people/F.Ciravegna/ATEM03/cimiano-ecml03-atem.pdf (состояние страницы на 14.09.2016).

14. Cimiano P., Hotho A., Staab S. Learning Concept Hierarchies from Text Corpora Using Formal Concept Analysis // Journal of Artificial Intelligence research. – 2005. – Vol. 24, № 1. – P. 305–339.

15. Kavalec M., Svatek V. A Study on Automated Relation Labelling in Ontology Learning // Ontology Learning and Population / eds. P. Buitelaar, Ph. Cimiano. – Amsterdam: IOS Press, 2005. – P. 44–58.

16. Bates M.J. Indexing and access for digital libraries and the Internet: human, database, and domain factors // Journal of the American Society for Information Science. – 1998. – Vol. 49. – P. 1185–1205.

17. Understanding Information Retrieval Systems: Management, Types, and Standards / ed. M.J. Bates. – Boca Raton: Taylor & Francis Group, LLC, 2012. – 752 p.

18. Borko H. Toward a theory of indexing // Information Processing & Management. – 1977. – Vol. 13, № 6. – P. 355–365.

19. Buckland M. Vocabulary as a central concept in library and information science // Digital libraries: interdisciplinary concepts, challenges, and opportunities. Proceedings of the Third international conference on conceptions of library and information science. CoLIS3. Dubrovnik. Croatia. 23-26 May 1999 / eds. T. Aparac, T. Saracevic, P. Ingwersen, P. Vakkari. – Lokve: Benja Publishing, 1999. – P. 3–12.

20. Candan K.S., Di Caro L., Sapino M.L. Creating tag hierarchies for effective navigation in social media // International Conference on Information and Knowledge Management, Proceedings. – Napa Valley, 2008. – P. 75–82.

21. Chan L.M. Inter-indexer consistency in subject cataloging // Information Technology and Libraries. – 1989. – Vol. 8, № 4. – P. 349–357.

22. Chen C. Mapping scientific frontiers: the quest for knowledge visualization. – Berlin: Springer, 2003. – 256 p.

23. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. – 1990. – Vol. 41, № 6. – P. 391–407.

24. Farrow J. All in the mind: concept analysis in indexing // The Indexer. – 1995. – Vol. 19, № 4. – P. 243–247.

25. Frické M. Logic and the Organization of Information: An Introduction. NASKO // North American Symposium on Knowledge Organization. – 2013. – P. 70–75. – URL: http://journals.lib.washington.edu/ index.php/nasko/article/view/14646 (состояние страницы на 14.09.2016).

26. Golub K. Automated subject classification of textual web documents // Journal of Documentation. – 2006. – Vol. 62, № 3. – P. 350–371.

27. Mazzocchi F. Ranganathan's universe of knowledge and categorical thinking // SRELS Journal of Information Management. – 2013. – Vol. 50, № 6. – P. 763–778.

28. Gnoli C. Ten long-term research questions in knowledge organization // Knowledge Organization. – 2008. – Vol. 35, № 2/3. – P. 137–149.

29. Ranganathan Sh.R. Colon classification. – London: Edward Goldston, 1933. – 106 p.

30. Broughton V. Brian Vickery and the Classification Research Group: the legacy of faceted classification // Facets of knowledge organization: proceedings of the ISKO UK Second Biennial Conference, 4-5 July 2011, London / eds. A. Gilchrist, J. Vernau. – London: Emerald, 2012. – P. 315–326.

31. Broughton V., Slavic A. Building a faceted classification for the humanities: principles and procedures // Journal of Documentation. – 2007. – Vol. 63, № 5. – P. 727–754.

32. Nelson D., Turney L. What's in a word? Rethinking facet headings in a discovery service // Information Technology and Libraries. – 2015. – Vol. 34, № 2. – P. 76–91.

33. Rose S., Roberts I., Cramer N. Facets for discovery and exploration in text collections // Workshop on Interactive Visual Text Analytics for Decision Making, 2011. – URL: http://vialab.science.uoit.ca/extvis2011/papers/textvis%202011-rose.pdf (состояние страницы на 14.09.2016).

34. Will L. Rigorous Facet Analysis as the Basis for Constructing Knowledge Organization Systems (KOS) of All Kinds // ISKO UK Conference. – London, 2013. – URL: http://www.iskouk.org/sites/default/files/WillPaper.pdf (состояние страницы на 14.09.2016).

35. Hjørland B. Facet analysis: The logical approach to knowledge organization // Information Processing and Management. – 2013. – Vol. 49, № 2. – P. 545–557.

36. Zufferey S., Degand L. Annotating the Meaning of Discourse Connectives in Multilingual Corpora // Corpus Linguistics and Linguistic Theory. Pages 1–24, ISSN (Online) 1613-7035, DOI: 10.1515/cllt-2013-0022, September 2013. – URL: http://www.academia.edu/download/32477556/Zufferey-DegandCLLT_2013-0033.pdf (состояние страницы на 14.09.2016).

37. Meyer T., Popescu-Belis A., Hajlaoui N., Gesmundo A. Machine Translation of Labeled Discourse Connectives // Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2012). – San Diego, CA, 2012. – URL: https://infoscience.epfl.ch/record/192524/files/Meyer_AMTA_2012.pdf (состояние страницы на 14.09.2016).

38. Kortmann B. Adverbial Subordination. A Typology and History of Adverbial Subordinators Based on European Languages. – Berlin/New York: Mouton de Gruyter, 1997. – 425 p.

39. Breindl E., Volodina A., Waßner U.H. Handbuch der deutschen Konnektoren 2. Semantik der deutschen Satzverknüpfer. 2 Teilbände. – Berlin/Boston: de Gruyter, 2014. – 1307 s.

40. Inkova O. Le relazioni logico-semantiche tra gli enunciati: una proposta di classificazione // In Ricerche in slavistica / eds. M. di Filippo, F. Esvan. – Naples: L'Orientale University Press (in press).

41. Инькова О.Ю. К проблеме описания многокомпонентных коннекторов русского языка: *не только… но и* // Вопросы языкознания. – 2016. – № 2. – С. 37–60.

42. Инькова-Манзотти О.Ю. Коннекторы противопоставления во французском и русском языках. Сопоставительное исследование. – М.: Информэлектро, 2001. – 434 с.

43. Зацман И.М., Инькова О.Ю., Кружков М.Г., Попкова Н.А. Представление кроссязыковых знаний о коннекторах в надкорпусных базах данных // Информатика и ее применения. – 2016. – Т. 10, № 1. – С. 106–118.

Материал поступил в редакцию 27.10.16.

Сведения об авторах

ЗАЦМАН Игорь Моисеевич – доктор технических наук, зав. отделом Института проблем информатики Федерального исследовательского центра «Информатика и управление» РАН (ФИЦ ИУ РАН), Москва
e-mail: izatsman@yandex.ru

ИНЬКОВА Ольга Юрьевна – доктор филологических наук, старший научный сотрудник Института проблем информатики ФИЦ ИУ РАН
e-mail: olyainkova@yandex.ru.

НУРИЕВ Виталий Александрович – кандидат филологических наук, старший научный сотрудник Института проблем информатики ФИЦ ИУ РАН, Института языкознания РАН, Москва
e-mail: nurieff.v@gmail.com.

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2 : 001.4

Н.А. Кочеткова, П.Д. Ермаков

Метод извлечения однословных терминов на основе статистического распределения слов внутри контекста

Представлен метод извлечения однословных терминов из монотематических коллекций документов, построенный на статистическом распределении терминов в контексте. Метод основан на подсчете дивергенции Кульбака–Лейблера относительно получаемого по контексту распределения Ципфа. Приводятся результаты экспериментов на коллекциях текстов журналов «Системы автоматизации проектирования и компьютерная графика» и «Вестник Томского Государственного университета. Биология». Подобранный материал позволил сделать выводы о пригодности предложенного метода для технической и естественнонаучной областей знаний. Предложенный метод в комбинации с существующими решениями позволяет существенно повысить количество терминов, автоматически извлекаемых из одной и той же коллекции текстов предметной области.

Ключевые слова: *автоматизированное извлечение терминов, меры терминологичности, извлечение знаний из текстов, однословные термины*

ВВЕДЕНИЕ

Корпуса текстов предметных областей и извлеченные из них данные имеют высокую прикладную ценность. Среди прочего, результаты анализа таких корпусов могут быть использованы при составлении терминологических словарей, классификаторов, рубрикаторов, при автоматическом индексировании и реферировании документов, автоматической классификации и кластеризации документов, в информационном поиске и машинном переводе. На основе корпусов предметных областей создаются и пополняются терминологические базы данных, тезаурусы, формальные онтологии для отдельных предметных областей, а также многоязычные терминологические ресурсы.

Многие проблемы, возникающие при анализе текстов специальных предметных областей, не имеют однозначных решений. К таким проблемам можно отнести вопросы о том, что считать термином той или иной предметной области, как составить корпус текстов предметной области, как выделить термины из текстов в таком корпусе и др.

В настоящем исследовании под термином мы понимаем лексическую единицу, характерную для некоторой коллекции текстов предметной области и являющуюся названием некоторого понятия из этой области. Термины существуют в рамках определённой терминологии, т. е. входят в конкретную лексическую систему языка через конкретную терминологическую систему.

Для выделения терминов обычно используются различные статистические методы, вычисляющие меры терминологичности, которые позволяют автоматически выделять кандидатов в термины из текстов, а затем ранжировать их по степени терминологичности в данной коллекции с помощью значений выбранных мер.

Особенно актуален статистический подход в случаях, когда изучаются процессы становления новой предметной или новой междисциплинарной области, изменения терминологии.

При извлечении терминов предметной области больше внимания уделяется извлечению терминологических словосочетаний, и значительно меньше исследований посвящено извлечению отдельных слов-терминов [1, 2].

Известно, что список самых частотных словосочетаний, извлеченных из текстов предметной области, содержит очень высокую долю терминологических словосочетаний (порядка 50%). Подавляющее же количество наиболее частотных слов, извлеченных из коллекции текстов предметной области, напротив, не является терминами. Спорным остается и

вопрос о частях речи однословных терминов. Например, в [3] к однословным терминам отнесены не только существительные, но и прилагательные, и глаголы. При этом отмечается, что подавляющее большинство терминов предметной области составляют именно существительные. В отличие от слов общего языка, термины слабее связаны с контекстом [4–6].

На данный момент существует много работ, посвященных автоматизированному извлечению терминов для различных языков. Так в [7–9] приведен анализ для английского, итальянского и французского языков, соответственно. Методы, наиболее эффективные для русского языка, и их подробная оценка приведены в работах [2, 10].

Стоит отметить, что меры терминологичности, основанные на частотных показателях, дают лучшие результаты на словосочетаниях, однословные же термины зачастую игнорируются. [11]. При том, что для технических и естественнонаучных текстов средняя длина термина составляет 2-3 слова, тогда как в таких областях как искусство превалируют однословные термины. В некоторых методах учитывалась зависимость средней длины от жанра [12].

Среди статистических методов, позволяющих оценить терминологичность однословных кандидатов, лучшие показатели точности дают меры Weirdness и C-value [2]. Однако уровень точности составляет порядка 60-80% (в зависимости от выбранных текстов предметной области, применяемых шаблонов и фильтров). Мера Weirdness требует наличия контрастной коллекции внушительного объема, которая заведомо не будет иметь пересечения с терминами исходной предметной области, но сможет отфильтровать присущие целевой коллекции стилистические обороты (таким образом термины высокого уровня абстракции, такие как *«алгоритм», «система», «признак объекта»*, будут исключены из рассмотрения вместе со стилевыми конструкциями). Мера C-value, штрафует короткие сочетания, входящие в состав более длинных и изначально была рассчитана на сочетания длиной более 2-х, представляющие собой определенный класс именных групп. Практика показывает, что размер штрафа, получаемый для униграмм, недостаточен для значительного изменения их упорядочивания.

Вопрос применимости подобных мер к тематическим техническим коллекциям до сих пор исследован слабо. Например, в работе [2] рассматривались новостные статьи и гуманитарные журналы, поэтому возникает вопрос об эффективности применения описанных в ней методов к техническим и естественнонаучным текстам.

Распространены и методы на основе вероятностных распределений коллокатов. К ним относятся меры, основанные на вычислении энтропии [13–16] и дивергенции Кульбака-Лейблера (использовалась, например, среди прочих мер в [17] для классификации сочетаний вида глагол+существительное).

В результате проведенных экспериментов [18] мы обнаружили, что биграммы каждого слова (с левым или правым контекстом, а также с двумя контекстами одновременно) распределены в соответствии с законом Ципфа, степенной показатель которого варьируется в широких пределах. При этом, те из слов, которые образуют двухсловный термин, имеют показатель степени почти в 2 раза больший, чем средний для всех слов этого корпуса.

Мы предлагаем новый метод выделения однословных терминов из научных текстов, достоинством которого является то, что для точного выделения однословных терминов не требуется дополнительная контрастная коллекция текстов.

ОПИСАНИЕ МЕТОДА

Для извлечения однословных терминов мы использовали результаты наших предыдущих экспериментов [18], которые показали, что если из текста после лемматизации извлечь все биграммы для фиксированного слова (неважно, в какой оно будет позиции в биграмме) и расположить их по убыванию частоты встречаемости, то полученное распределение подчиняется закону Ципфа. В настоящей работе мы предположили, что для слов, не являющихся терминами данной предметной области, степенной показатель распределения Ципфа будет близок к 1. Для одиночных слов в работе [19] для русского языка было получено среднее значение степенного коэффициента, равное 0,97. Для биграмм, по нашим данным, подобный показатель равен примерно 0,9. Наше предположение состоит в том, что для терминов этот показатель должен существенно отличаться в бóльшую сторону.

Согласно работам [4, 5], контекст термина отличается низкой энтропией. Это связано с тем, что он имеет довольно ограниченный набор встречающихся в нем слов, тогда как «обычное» слово будет требовать разнообразных уточнений, не выявляя при этом связываемых коллокатов. Исключением могут быть термины, входящие в другие сочетания. Подобные ситуации должны фильтроваться заранее или по результатам экспериментов.

Таким образом, предлагаемый нами метод выделения однословных терминов из научных текстов заключается в следующем.

Из монотематической размеченной коллекции с лемматизацией, частеречной разметкой и снятой омонимией извлекаются все биграммы. Помимо этого, для коллекции составляется словарь начальных форм. Для всех начальных форм, встречаемость которых выше заданного порога, проводятся следующие действия: для всех биграмм, в которых встречается выбранное слово, рассчитывается частота их встречаемости, затем полученный список биграмм длины n сортируется по убыванию этой частоты, далее берется биграмма с максимальной частотой встречаемости P_1 и на ее основе рассчитывается распределение Ципфа с таким же количеством элементов: $Q_i = P_1/i$, i=1...n. Для этих двух распределений рассчитывается значение дивергенции Кульбака–Лейблера: $D(P||Q) = \sum(P_i*\log(P_i/Q_i))$, i=1...n. Слова с максимальным значением дивергенции считаются кандидатами в термины. Они должны быть просмотрены экспертами для окончательного принятия решения.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для извлечения слов-терминов мы использовали коллекции статей журналов «Системы автоматизации проектирования и компьютерная графика» (7,2 млн словоупотреблений) и «Вестник Томского Государственного университета. Биология» (1,8 млн словоупотреблений). Такой выбор материала позволил сделать выводы о применимости предлагаемого метода как к техническим, так и к естественнонаучным текстам.

Далее была проведена лемматизация и снятие омонимии [20], выделены все существительные, для каждого из них получены списки его биграмм, которые извлекались лишь в рамках одного предложения; концом предложения считался знак «точка». При этом учитывались сочетания со знаками препинания, так как такой контекст показателен.

Затем для каждого существительного, частота встречаемости которого превысила порог в 10 употреблений на коллекцию, по описанному алгоритму было подсчитана дивергенция Кульбака–Лейблера:

$$\mathrm{Div}(P, Q) = \sum(P_i * \log(P_i * i / P_1)),$$

где: i=1..n ;

P_i – вероятность *i*-го элемента в списке биграмм длины *n*.

Для сравнения результатов были выбраны 50 слов с максимальным значением для следующих мер:

- частота в коллекции;
- C-value [12]

C-value(a)= log|a|*freq(a), если не вложен
log|a|*freq(a)-1/N(Ta)*∑freq(b), в противном случае,

где: *a* – кандидат в термины; |a| – длина словосочетания, измеряемая количеством слов; freq(*a*) – частотность *a*; *Ta* – множество словосочетаний, которые содержат *a*; N(*Ta*) – количество словосочетаний, содержащих *a*; ∑freq(*b*) – сумма частот всех сочетаний, содержащих *a*; *b* – сочетания, содержащие *a*;

- странность (Weirdness) [21]

Weirdness = (Ws/Ts)/(Wg/Tg) ,

где: Ws – частотность слова в коллекции предметной области; Ts – общее количество словоупотреблений в коллекции предметной области; g – контрастная коллекция; Wg – частотность слова в контрастной коллекции; Tg – общее количество словоупотреблений в контрастной коллекции;

- SupW(w,g) = (Ws*Ws/Ts)/((Wg+Ws)/(Tg+Ts)) [22].

В качестве контрастной коллекции при подсчете Weirdness и SupW была использована коллекция произведений «Детективы» (15,8 млн словоупотреблений). Оценка точности результатов проводилась экспертом.

Для топ-50 терминов показатели всех мер являются высокими (табл. 1). Это может быть объяснено тем, что общенаучные понятия при данной оценке были отнесены к правильным ответам. Однако стоит отметить, что списки, полученные для C-value, Weirdness и SupW, по своему составу очень мало отличаются от списка, полученного для частоты, и друг от друга, в то время как дивергенция в большинстве случаев выявляет другие термины. Это связано с тем, что предлагаемая нами мера терминологичности не зависит непосредственно от частоты самого слова, а зависит от частотного распределения его сочетаний. В результате, максимальные значения дивергенции получили более специфичные термины, которых нет в списках, полученных по частоте (табл. 2 и 3). Стоит отметить, что во всех полученных списках не наблюдается упорядочивание по терминологичности: нельзя сказать, что слово, получившее максимальное значение, – самый характерный термин предметной области.

Списки топ-100 терминов, полученные в ходе экспериментов, характеризуют выбранные коллекции скорее как коллекции-сборники с большим разнообразием тем внутри предметной области. В терминах коллекции журнала «САПР и компьютерная графика» отражены обе заявленные темы. В коллекции журнала «Биология» наблюдается большое разнообразие тем, при этом выделяются общие для подобластей географические и классификационные термины.

Можно сделать вывод о высокой общности тематики рассматриваемых журналов, с одной стороны, и о том, что степень тематической однородности коллекции научных текстов соотносится с однородностью множества выделяемых терминов – с другой. Полученные нами списки терминов дают представление об обобщенном содержании коллекций. Этих списков достаточно, чтобы получить предварительную информацию о наиболее важных объектах исследования, материале и методах.

Таблица 1

Доля верных ответов в топ-50 значениях каждой из мер, %

Коллекция	Частота	C-value	Weirdness	SupW	Дивергенция
«Биология»	86	90	76	94	92
«САПР и компьютерная графика»	90	90	92	98	90

Таблица 2

Топ-100 терминов коллекции «Биология»

Частота	SupW	Дивергенция
вид, почва, растение, метр, исследование, условие, сантиметр, содержание, часть, состав, процесс, результат, тип, данные, лес, развитие, наука, уровень, анализ, район, территория, область, материал, среда, рост, сообщество, влияние, изменение, участок, количество, период, состояние, система, число, показатель, метод, форма, структура, томск, вода, дерево, университет, активность, горизонт, значение, зона, фактор, работа, покров, клетка, кедр, особь, особенность, изучение, сравнение, степень, ярус, побег, концентрация, вещество, связь, высота, раз, оценка, площадь, воздействие, культура, элемент, животное, использование, хромосома, россия, популяция, растительность, формирование, признак, численность, порода, флора, склон, длина, биология, доля, увеличение, основа, действие, лист, масса, образ, характеристика, озеро, сосна, профиль, институт, ряд, предел, течение, возраст, распределение, качество	почва, растение, исследование, сантиметр, содержание, состав, метр, сообщество, вид, наука, развитие, условие, показатель, анализ, структура, активность, процесс, кедр, данные, особь, покров, концентрация, ярус, территория, фактор, хромосома, популяция, уровень, формирование, лес, численность, среда, зона, метод, изучение, растительность, вещество, биология, горизонт, культура, увеличение, материал, побег, распределение, элемент, подрост, система, древостой, воздействие, характеристика, ареал, район, тайга, насаждение, порода, экология, экосистема, семя, гумус, клетка, флора, профиль, коэффициент, сравнение, днк, рост, сосна, методика, местообитание, кедровник, плотность, основа, сеянец, число, фракция, накопление, распространение, тип, температура, кислота, параметр, прирост, сосняк, микрометр, гриб, азот, устойчивость, белок, ландшафт, участок, разнообразие, полевка, бактерия, животное, калий, степь, обработка, спектр, самка, биомасса	ген, лист, уровень, показатель, участок, метод, анализ, форма, метр, орган, развитие, активность, система, род, территория, данные, компонент, образ, сантиметр, особь, структура, период, сообщество, клетка, семя, дерево, кедр, факт, формирование, состояние, рост, зона, популяция, цвет, концентрация, комплекс, площадь, распределение, возраст, класс, растительность, горизонт, численность, болото, масса, спектр, белок, продукт, раствор, побег, насаждение, животное, мех, местообитание, вода, объем, озеро, боль, плод, семейство, соотношение, древостой, корень, реакция, механизм, фрагмент, препарат, отсутствие, накопление, определение, поверхность, ландшафт, слой, хром, функция, фракция, стадия, разнообразие, гриб, мед, ареал, ярус, глубина, склон, объект, центр, качество, цель, представитель, тон, индекс, цикл, выборка, интенсивность, этап, отношение, запас, мощность, флора, сход

Таблица 3

Топ-100 терминов коллекции «САПР и компьютерная графика»

Частота	SupW	Дивергенция
система, работа, модель, данные, проектирование, проект, изделие, компания, элемент, пользователь, процесс, деталь, программа, предприятие, решение, объект, обработка, чертеж, задача, документ, технология, расчет, информация, производство, инструмент, модуль, разработка, сапр, изменение, управление, параметр, продукт, поверхность, материал, качество, версия, моделирование, документация, файл, оборудование, схема, конструкция, специалист, внедрение, обеспечение, функция, форма, анализ, станок, этап, операция, размер, режим, применение, комплекс, сборка, команда, построение, формат, автоматизация, метод, приложение, организация, образ, требование, план, линия, спецификация, окно,	система, модель, данные, проектирование, работа, проект, изделие, элемент, пользователь, программа, предприятие, процесс, деталь, объект, компания, обработка, чертеж, технология, задача, производство, расчет, модуль, инструмент, разработка, документ, сапр, параметр, продукт, информация, управление, моделирование, материал, файл, документация, поверхность, внедрение, оборудование, конструкция, схема, функция, станок, этап, анализ, специалист, сборка, режим, построение, комплекс, формат, автоматизация, приложение, спецификация, конструктор, заказчик, таблица, разработчик, интерфейс, требование, операция, чпу, редактирование, геометрия, стандарт, изображение, библиотека, объем, линия, печать, переход, поддержка, профиль, цикл, проектировщик, оснастка, интеграция, трубопровод, механизм, доступ, продукция, узел, технолог, кон-	процедура, фрагмент, алгоритм, тон, аппарат, текст, код, граница, компонента, металл, видео, конфигурация, серия, порт, инженер, курс, представление, масштаб, действие, провод, гост, длина, мастер, отображение, перемещение, панель, сила, представитель, тест, клиент, шаблон, поле, дизайн, реализация, опция, карта, стратегия, каталог, функциональность, сетка, эскиз, компьютер, метр, лист, перечень, фото, подсистема, пресс, стоимость, грань, бизнес, техник, движение, сцена, шаг, соединение, высота, пласт, концепция, путь, принцип, позиция, методика, машина, координата, обновление, кривая, корпус, преобразование, фон, ось, траектория, монитор, распреде-

Частота	SupW	Дивергенция
библиотека, структура, конструктор, условие, связь, заказчик, группа, таблица, значение, состав, разработчик, интерфейс, изготовление, задание, изображение, соответствие, настройка, формирование, чпу, развитие, среда, редактирование, геометрия, стандарт, поддержка, печать, переход, объем, направление, узел, контроль	троль, ось, установка, траектория, архив, производительность, связь, блок, реализация, контур, нагрузка, подразделение, шаблон, комплект, эксплуатация, затрата, сечение, гост, определение	ление, поиск, свет, взаимодействие, показатель, правило, вывод, размещение, здание, сканер, сеть, интеграция, диск, копия, последовательность, ресурс, отчет, поставщик, канал, отверстие, визуализация, пик, диаметр, расширение, справочник, сопряжение, загрузка

Таблица 4

Количество совпавших терминов в топ-100 для разных пар мер

Сравниваемые меры	«Биология»	«САПР и компьютерная графика»
Частота и SupW	57	73
Частота и Div	41	0
SupW и Div	28	4

Представление об обобщенном содержании коллекции журнала «САПР и компьютерная графика» будет более полным, чем о коллекции журнала «Вестник Томского Государственного университета. Биология».

Мы подсчитали количество слов, находящихся одновременно в двух списках, полученных с использованием различных мер (табл. 4).

Столь небольшая доля совпадений для дивергенции показывает, что предлагаемая мера терминологичности выделяет термины другой природы. А так как в последнее время во многих работах ([2] вслед за [23] и [1]) показано, что комбинирование признаков для выделения терминов позволяет значительно улучшить качество выделения даже при использовании схожих по своей природе мер, то добавление в композицию меры, использующей другие данные, должно дать существенное улучшение показателей точности и полноты метода.

ЗАКЛЮЧЕНИЕ

Нами предложен новый метод выделения однословных терминов из научных текстов, основанный не на частоте слова, а на частотном распределении его сочетаний. Несмотря на то, что это исследование можно считать сугубо предварительным, основные гипотезы в ходе экспериментов на рассматриваемом материале подтвердились.

Предлагаемый метод может быть успешно использован для выделения одиночных терминов из коллекции текстов предметной области. Результаты экспериментов показали, что по точности предлагаемый метод не уступает другим широко известным статистическим методам выделения терминов, а также, что он позволяет выделять термины, не выявляемые на основе частоты употребления слова, а основываясь на ином принципе определения терминологичности.

Эксперименты показали, что тематическая однородность выделяемых терминов характеризует степень тематической однородности коллекции.

К недостаткам метода стоит отнести его неустойчивость к таким ошибкам морфологического анализатора, как неверное определение части речи и леммы, и неточности графематического анализа. Если не ограничиваться первой сотней выделяемых методом кандидатов в термины, а проанализировать первую тысячу, то в ней можно будет найти некоторые местоимения, а также обрывки слов, получаемые при переносах.

СПИСОК ЛИТЕРАТУРЫ

1. Zhang Z., Iria J., Brewster Ch., Ciravegna F. A Comparative Evaluation of Term Recognition Algorithms // In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. – P. 2108-2113.
2. Лукашевич Н.В., Логачев Ю.М. Использование методов машинного обучения для извлечения слов-терминов // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием – КИИ-2010, г. Тверь, 2010. – URL: http://www.raai.org/resurs/papers/kii-2010/
3. Митрофанова О.А., Захаров В.П. Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». – М.: Изд-во РГГУ, 2009. – С. 321-328.

4. Manin D. Zipf's Law and Avoidance of Excessive Synonymy // Cognitive Science. – 2008. – Vol. 32, Iss. 7. – P. 1075–1098. – URL: http://onlinelibrary.wiley.com/doi/10.1080/03640210802020003/epdf
5. Ferrer i Cancho R. Hidden communication aspects in the exponent of Zipf's law // Glottometrics. – 2005. – P. 98–119. – URL: http://groups.lis.illinois.edu/amag/langev/paper/ferrer05hiddenCommunicationZipf.html
6. Маслов В.П., Маслова Т.В. О законе Ципфа и ранговых распределениях в лингвистике и семиотике // Математические заметки. – Т. 80, вып.5. – М. : Наука, 2006. – С. 718-732.
7. Merkel Magnus, Foo Jody, Ahrenberg Lars. Iphraxtor – A linguistically informed system for extraction of term candidates // In Proceedings of the 19th Nordic Conference on Computational Linguistics (Nodalida 2013), May 22-24. – Oslo, 2013. – P. 121-132.
8. Bonin Francesca, Dell'Orletta Felice, Venturi Giulia, Montemagni Simonetta. A Contrastive Approach to Multi–word Term Extraction from Domain Corpora // In Proceedings of the LREC 2010, Seventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA). – Valletta, Malta, 2010. – P. 3222-3229.
9. Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases // In Proceedings of the COLING-92. – Nantes, France, 1992. – P. 977–981.
10. Браславский П.И., Соколов Е.А. Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». – М..: Изд-во РГГУ, 2008. – С. 67-74.
11. Ягунова Е.В., Пивоварова Л.М. От коллокаций к конструкциям // Acta Linguistica Petropolitana. Труды института лингвистических исследований РАН. – СПб., 2011. – С. 568-617.
12. Frantzi K., Ananiadou S., Mima H. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method // International Journal of Digital Libraries. – 2000. – Vol. 3(2). – P. 117-132.
13. Shimohata Sayori, Sugio Toshiyuki, Nagata Junji. Retrieving collocations by co-occurrences and word order constraints // In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics. – 1997. – P. 476–481.
14. Resnik Philip. Selectional preference and sense disambiguation // In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How. – 1997. – P. 52–57. – URL: http://www.aclweb.org/anthology/W97-0209
15. Sag Ivan, Baldwin Timothy, Bond Francis, Copestake Ann, Flickinger Dan. Multiword expressions: A pain in the neck for NLP // Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, 2002. – P. 189–206.
16. Ramisch Carlos, Schreiner Paulo, Idiart Marco, Villavicencio Aline. An evaluation of methods for the extraction of multiword expressions // In Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions, European Language Resources Association (ELRA). – Marrakech, Morocco, 2008. – P. 50–53.
17. Fazly Afsaneh, Stevenson Suzanne. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measure // In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Association for Computational Linguistics. – 2007. – P. 9–16.
18. Кочеткова Н.А., Клышинский Э.С., Ермаков П.Д. Подчиняются ли составные конструкции закону Ципфа? // Системный администратор. – 2016. – № 11. (Принята в печать).
19. Gelbukh A., Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language // Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science. – 2001. – № 2004. – P. 332-335.
20. Рысаков С.В. Методы борьбы с омонимией // Системный администратор. – 2015. – №10. – С. 92-95.
21. Ahmad K., Gillam L., Tostevin L. University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval // In the Proceedings of Eighth Text Retrieval Conference (Trec-8), 1999. – URL: http://trec.nist.gov/pubs/trec8/t8_proceedings.html
22. Кочеткова Н.А. Метод извлечения технических терминов с использованием усовершенствованной меры странности // Научно-техническая информация Сер. 2. – 2015. –№ 5. – С. 25-32.
23. Pecina P., Schlesinger P. Combining association measures for collocation extraction // In Proceedings of the 21th international conference on computational linguistics and 44th annual meeting of the association for computational linguistics (COLING/ACL 2006). – Sydney, Australia, 2006. – P. 651-658.

Материал поступил в редакцию 29.10.16.

Сведения об авторах

КОЧЕТКОВА Наталия Александровна – аспирант, Департамент компьютерной инженерии, Московский институт электроники и математики, Научно-исследовательский университет «Высшая Школа Экономики» (ДКИ МИЭМ НИУ ВШЭ)
e-mail: natalia_k_11@mail.ru

ЕРМАКОВ Петр Дмитриевич – аспирант ДКИ МИЭМ НИУ ВШЭ
e-mail: ermakov.pd@gmail.com

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК [004.5 : 81’322] : 004.738.5

Е.Н. Пименов

Правила подготовки запросов в лингвистическом освещении

Исследуются рекомендации по составлению запросов в Интернете и, в частности, правила подготовки запросов для комплектования сетевого ресурса - путеводителя по сохранности документов. К указанным рекомендациям приводится лингвистический комментарий.

Ключевые слова: *правила, составление запросов, Интернет, типы запросов, запросы обиходно-бытового назначения, запросы специального назначения*

ВВЕДЕНИЕ

Правила составления запросов в сети имеют характер практических рекомендаций без четко очерченных оснований и обосновываются главным образом тем, что использование определенного правила служит для повышения качества поиска. В настоящей работе показана преимущественно лингвистическая сторона подготовки запросов, в свете которой описываемые правила получают свою более адекватную характеристику.

По общему назначению сетевые запросы разделяются на две разновидности:

- запросы на получение из Интернета справочной информации бытового характера, как, например, *Какая погода будет в …?, Где купить …. ?, Как проехать в … ?* Значение этих запросов соответствует «обиходным» информационным потребностям и общезначимым семантическим категориям, таким как *погода, торговля, транспорт, учреждения, здоровье, животные, кухня, компьютеры, развлечения* и т. п. Актуальность тематики делает такие запросы часто употребимыми (массовыми) в Интернете. К специфике этих запросов относится то, что они составляются на обычном естественном языке;
- запросы, имеющие не обиходно-бытовое, а специальное назначение. Поскольку запросов с бытовым содержанием в Интернете имеется заведомо больше, чем направленных, например, на решение научных задач, такие запросы в сети являются относительно редкими. Язык составления этих запросов имеет характер достаточно сложного и искусственного языкового конструкта.

Правила подготовки запросов, описываемые в настоящей статье, относятся главным образом ко второму типу (запросам специального назначения). Что касается «обиходных» запросов, то они составляются так же, как обычные сообщения чего-то о чем-то, а не с применением рассматриваемых правил. Собственно подготовки запросов как формулирования информационной потребности пользователя ключевыми словами специального информационно-поискового языка (перевода на данный язык) при таком выражении информационных потребностей не происходит. Думается, что многие лица, активно работающие с Интернетом, не имеют понятия о каких-либо правилах составления запросов, языке их подготовки и тому подобных вещах. Необходимость в этих понятиях возникает тогда, когда речь идет о специальных информационных запросах.

СУЩЕСТВУЮЩИЕ РЕКОМЕНДАЦИИ ПО СОСТАВЛЕНИЮ ЗАПРОСОВ

Имеющиеся рекомендации по подготовке запросов в сети (их было рассмотрено не менее 60) можно свести к семи основным, направленным на улучшение качества поиска, т.е. параметров точности или полноты результатов информационного поиска.

Для повышения точности поиска в Интернете рекомендуются следующие правила.

1. *Ставить узкие по содержания запросы, конкретизируя их содержание ключевыми словами*. Словесные формулировки данного правила: “Правильный запрос состоит из нескольких слов, потому что по одному слову обычно трудно понять, что вы хотите найти” [1]; “Для достижения необходимого результата уточняйте (конкретизируйте) запрос, используя ключевые слова. Если вы ищите сведения по автомобилю Toyota, например, то ваш вопрос должен быть: "автомобиль toyota", а не "автомобиль"” [2];

“Для того, чтобы поисковая выдача наиболее полно соответствовала вашему запросу, необходимо как можно подробнее уточнить свой вопрос” [3].

2. *Конкретизировать содержание запроса с использованием команд-операторов,* задающих условия поиска в Интернете [4-11]. В Google такими являются операторы:

& – конъюнкция, одновременное присутствие в запросе двух или более терминов

“” (кавычки) – поиск определенной фразы дословно

«-» (минус) – исключение слова при поиске

***** (символ маски) – условное обозначение произвольного количества любых символов

.. (числовой интервал) – выражение 2000..2005 означает все года от 2000 до 2005

Info: – поиск информации об указанном сайте. Пример: info:webostrovok.ru

Site: – поиск на определенном сайте или домене

Link: – поиск страниц, которые ссылаются на заданный сайт

Cache: – поиск в кэше Google

Intitle: – поиск по заголовкам страницы

Inurl: – поиск по URL

Intext: – при поиске не учитываются заголовки страниц и просматривается только текст тела страницы

Inanchor: – поиск в тексте якоря страницы

Filetype: – поиск в файлах с заданным расширением

Define: – оператор для быстрого поиска определений

Movie: – поиск названий фильмов, как movie: One Flew Over the Cuckoo’s Nest

Music: – поиск информации о музыке

3. *Использовать подсказки при наборе запросов*: “Если вы не уверены, как сформулировать запрос, система подсказок поможет вам найти нужные слова” [12]; “Когда вы набираете в браузере свой поисковый запрос, вам по мере набора показываются различные варианты часто встречающихся запросов, вы можете выбрать один из них или ввести свой вариант” [13]. Так, при наборе запроса “Подсказки ключевых слов в Интернет” под поисковой формой Google появляются три варианта запроса: *Подсказки ключевых слов в You Tube*, *Подсказки ключевых слов в Google Adwords* и *Подсказки ключевых слов в Яндекс*. Поскольку при выборе одного из подсказываемых вариантов производится уточнение запроса, рассматриваемое правило следует отнести к косвенным средствам сужения содержания запроса.

Для повышения полноты результатов информационного поиска рекомендуются следующие правила.

4. *Расширить запрос с применением операторов.* С этой целью чаще всего употребляется операция дизъюнкции («или»), обозначаемая символами «OR» или «|». Данная операция соединяет слова, имеющие близкое содержание, в первую очередь – синонимы. Формулировки данного правила: “При поиске часто приходится переформулировать первоначальный запрос и вручную заменять слова на синонимы” [14]; “Если найденных страниц слишком мало или не имеется полезных страниц, попробуйте задать для поиска три-четыре синонима сразу” [15]. Ранее в Google применялся оператор синонимии ~ (тильда), служивший для расширения содержания запросов и подключавший к поиску синонимы слова, перед которым он расположен. В настоящее время от него отказались, и в основанном потому, что оператором пользовалось немного людей, а затраты времени и денег на его поддержку были значительными. При отказе от этого оператора «освободилось много ресурсов, которые можно использовать на другие, более продвинутые функции» [16].

5. *Использовать наиболее характерные для тематической области ключевые слова.* Словесные формулировки этого правила: “Вместо "сколько людей живёт в России" введите "население России", потому что именно этот термин будет использоваться на странице, посвящённой демографии. Запрос "аспирин" более информативен, чем "ацетилсалициловая кислота"” [17]. При поиске информации о писателе Вячеславе Рыбакове лучший результат оказался при добавлении в запрос слова «критика». “Идея заключается в том, чтобы при поиске использовать ключевые слова, характерные для сайтов того типа, который вы ищете. Например, странички хакеров можно обнаружить по характерному сленгу. Скажем, по ключевым словам *прога, пароль, юзер*” [18]. ”Даже если вы используете правильное слово, но большинство людей его не использует, слова может не оказаться на нужной странице. Например, запрос *Популярные рингтоны* более информативен, чем запрос *Популярные мелодии* ” [19].

6. *Использовать простую форму запроса*: “Опишите, что вам нужно, используя как можно меньше слов. Например, простой запрос [Погода Минск] даст лучшие результаты, чем более длинный запрос [Прогноз погоды для Минска Беларусь]” [20]; “Запросы должны быть простыми. Если вы ищете какое-то предприятие, просто введите его название или хотя бы ту часть названия, которую вы помните наверняка. Для большинства запросов вовсе не нужны редкие операторы или изощренный синтаксис. Чем проще, тем лучше” [21].

7. *Употреблять информативные ключевые слова*. “Используйте информативные слова, старайтесь избегать общих фраз и малоупотребительных синонимов” [22]. “Для поиска лучше использовать более информативные слова. Это увеличивает вероятность, что результаты будут релевантными. Такие слова, как *документ*, *веб-сайт*, *компания* или *информация*, обычно лишние” [23].

Две последние рекомендации представляют собой самые слабые из рассматриваемых правил и особенно то, что в запросах должны содержаться только информативные ключевые слова. Это требование очевидно и, вероятно, нет смысла его понимать как особое правило составления информационных запросов. То, что запросы должны быть простыми, в известном смысле вступает в противоречие с правилом 1, согласно которому однозначные по содержанию запросы должны обладать известной степенью синтаксической распространенности и, таким образом, сложности. Как и правило 5 (использование употребительной формы запроса), эта рекомендация дейст-

вует только по отношению к массовым и обиходно-бытовым информационным запросам, для составления которых не требуются редкие операторы или изощренный синтаксис. Что касается подготовки запросов, используемых для комплектования информационных ресурсов, то такие запросы в силу их назначения должны быть достаточно сложными. По указанным причинам рекомендации к правилам 6 и 7 можно далее в этой статье не рассматривать.

КОММЕНТАРИЙ К ОПИСАННЫМ ПРАВИЛАМ ПОДГОТОВКИ ЗАПРОСОВ

В комментарии к правилам подготовки запросов описывается применение рассматриваемых правил при создании путеводителя по сохранности документов. Путеводитель доступен по электронному адресу [24].

1. Сужение содержания запроса ключевыми словами

Согласно этому правилу, лучше строить запросы с большим количеством поисковых терминов. На качество поиска информации влияет не количество слов, составляющих информационный запрос, а то, что каждое слово в запросе сужает его содержание, узкие по содержанию запросы являются наиболее результативными. В некоторых случаях узкое содержание имеется даже в запросах, состоящих из одного, но очень конкретного слова. Такими являются, например, ключевые слова *Тетрациклин, Ацидобиофилин, Нескафе, Мерседес,* или слова, соответствующие одному означающему (денотату): *Росинант, Буцефал, Аполлон-Союз*, «Аврора», «Варяг» и др. Будучи поисковыми выражениями, эти слова не нуждаются в синтаксическом распространении для их уточнения, а поиск по ним не ведет к аномально большому информационному шуму.

Для поиска информации по другим (не очень узким) словам необходимо сужение их содержания. Повышение точности поиска при сужении запроса достигается тем, что многозначное слово в запросе ставится в определенный контекст, контекстным путем уточняется его содержание и устраняется многозначность. Таково лингвистическое основание рассматриваемого правила, одинаково действующего как в естественном языке, так и в случае составления запросов. Примеры контекстного устранения неоднозначности слова. Поиск по слову *Mercury* в контексте *Mercury Chemistry* дает информацию о ртути как о химическом элементе, а в контексте *Mercury Singer* – информацию о британском певце *Фредди Меркьюри*. Неоднозначность значения слова “Бор” устраняют контексты – *Сосновый бор*, *Бор элемент, химия* и *Нильс Бор* (датский физик-теоретик). Указанным образом разрешается омонимия поисковых терминов, являющаяся относительно редким явлением при составлении информационных запросов. Гораздо чаще, чем омонимия, таким же (контекстным) путем разрешается также неполное соответствие запросов информационным потребностям пользователя. Последний вид информации относится к промежуточной области между релевантной и шумовой информацией, когда о ней говорят, например, что это шум, но «не вредный» или, что рассматриваемая информация, в общем, похожа на нужную, но ею все же не является.

Неполное соответствие запросов информационным потребностям пользователя близко стоит к отношению полисемии, которое связывает близкие, но не полностью эквивалентные по значению слова. Поиск по полисемичным словам очень часто дает информацию, не требующуюся пользователю, но относящуюся к смежным тематическим областям. Например, в ответ на запросы, касающиеся биологической стороны консервации документов (*плесневые грибы, грибостойкость, дезинфекция, пестициды, биоциды* и многие другие слова), получаемая из сети информация чаще всего относится не к сохранности документов, а к сельскохозяйственной, медико-биологической, строительной и другой проблематике. Для «привязки» запроса к тематической области ‘консервация документов’ необходима специальная обработка запросов с применением их лексического уточнения.

Хорошей привязкой запросов к рассматриваемой тематической области являются ключевые слова *документы, книги, библиотеки, музеи, архивы*. Три последних термина называют хранилища документов, имплицитно содержат в себе содержание ‘сохранение’ и ‘документы’ (‘культурные ценности’) и таким образом вычленяют предметную область. В указанном контексте запрос “Консервация документов” представляется в виде серии подзапросов, включающей названные ключевые слова, как

Консервация документов
Консервация книг
Консервация в библиотеках
Консервация в архивах
Консервация в музеях
Сохранность документов
Сохранность книг
Сохранность библиотек
Сохранность архивов
Сохранность музеев
Защита …
……..

Поиск по приведенным запросам снимает возможную неоднозначность слов *консервация* и *сохранность*, и таким же путем производится уточнение большей части запросов, по которым проводится комплектование путеводителя по сохранности документов. Преимущество такой техники подготовки запросов заключается в том, что она обеспечивает высокие показатели точности и полноты результатов информационного поиска.

2. Сужение запроса с использованием команд-операторов

В ходе создания путеводителя по сохранности документов использовались три оператора Google: кавычки (“”), логическое отрицание (-) и оператор дизъюнкции (OR). Первые два оператора обладают способностью устранять многозначное содержание запросов.

В отношении функции уточнения запросов примечательна функция кавычек, с употреблением которых проводится поиск цитат, т.е. точного употребления поисковых терминов без учета их словообразовательных изменений и возможных синонимов. Запросы в кавычках приобретают значение, сравнимое со значе-

нием имен собственных, называющих конкретные, индивидуальные вещи, а не классы предметов, как нарицательные имена. Поиск по терминам, заключенным в кавычки, часто имеет высокую точность, и хорошие результаты дает, например, проведение поиска по названиям книг, песен, фильмов и географическим наименованиям объектов. Поиск по заключенным в кавычки личным именам является менее точным, последнее связано с тем, что словесное выражение личных имен может варьироваться в широких пределах.

Кроме полного наименования личного имени, как *Иван Степанович Иванов*, имеются еще формы с инициалами в препозиции и постпозиции к фамилии: *И.С. Иванов, Иванов И.С.* Последние формы являются наиболее употребительными в Интернете и поэтому более предпочтительными при проведении поиска. Так как за инициалами могут стоять самые разные имена и за фамилией *Иванов* – много разных персон, данные формы имен отличаются исключительной неоднозначностью. Если при поиске по фамилии автора необходимы высокая точность и полнота (когда, например, по какой-то фамилии составляется список научных трудов), приходится строить достаточно сложный информационный запрос с применением кавычек, операции дизъюнкции (ИЛИ) и операции логического отрицания. Так, для автора настоящий статьи данный запрос индексируется в следующем виде: *(«е н пименов» OR «пименов е н») «– пименова»*, где слово с логическим отрицанием «–Пименова» (без публикаций Пименовой) можно рассматривать как разновидность контекста, уточняющего и сужающего содержание запроса. Степень идентификации «определенного Пименова» и, соответственно, точность информационного поиска можно усилить, еще более уточняя запрос. Для этого можно использовать ключевые слова, называющие область научных интересов лица и характерные для его публикаций слова. С учетом последнего уточнения получится следующий информационный запрос: («е н пименов» OR «пименов е н») «– пименова» (индексирование OR тезаурусы OR путеводители). Интересно отметить, что в приведенном запросе одновременно представлено и сужение, и расширение содержания запроса. Эта «разнонаправленность» смысловой обработки одного и того же запроса является характерной чертой описываемой техники составления запросов.

3. Сужение запроса с использованием поисковых подсказок

В отличие от команд-операторов и лексических средств, используемых для сужения запросов, подсказки не прямо и непосредственно сужают запрос, но косвенным образом способствуют этому, показывая возможные пути уточнения содержания запроса. Главным назначением подсказок является сокращение времени набора запросов, когда тексты не набираются на клавиатуре, а выбираются из уже имеющихся. Вторым назначением подсказок является раскрытие возможного содержания слова путем описания типичных контекстов, устраняющих многозначность лексической единицы. Система дает варианты возможного содержания слова, как, например, *Наполеон* – **подсказки:** *Наполеон торт, Наполеон динамит* [фильм], *Наполеон Бонапарт, Наполеон Хилл* [американский писатель]. *Цезарь* – **подсказки:** *Цезарь сателлит* [охранная система], *Цезарь сателлит отзывы, Цезарь, Цезарь салат*. Выбор определенного варианта и означает сужение и уточнение запроса в соответствии с правилом 1 (сужение запроса ключевыми словами). Сетевые подсказки ориентированы на наиболее частые области применения Интернета, которыми, как известно, являются развлечения, информация о видных деятелях, кухне, интернет-бизнесе и других общезначимых областях. Эти подсказки не применимы по отношению к узким специализированным областям, и при создании путеводителя не использовались.

Следующие правила связаны с расширением содержания запросов.

4. Расширение с употреблением эквивалентных по содержанию слов. Равноценными или близкими по содержанию могут являться:

- синонимы. Английский термин *Acid Free Cardboard* (бескислотный картон) имеет синонимами ключевые слова *Acid Free Carton, Acid-Free Carton, Alkaline Cardboard, Alkaline Carton, Archival Cardboard, Archival Carton, Archival Quality Cardboard, Archival Quality Carton, Long Lived Cardboard, Long Lived Carton, Neutral Board, Neutral Cardboard, Neutral Carton, Permanent Cardboard, Permanent Carton.* В путеводителе информация о бескислотном картоне отбиралась по всем вышеназванным терминам. Наименование *Anoxic Treatment* (бескислотная обработка) эквивалентно по содержанию словам *Anoxic Display, Anoxic Fumigation, Anoxic Storage, Insect Control by Anoxia, Insect Control by Argon, Insect Control by Inert Gases, Low oxygen, Modified Atmosphere Treatment insect, Nitrogen Anoxia, Nitrogen Anoxia, Nitrogen treatment, Nitrogen treatment. Oxygen reduction, Oxygen-Free Environment*. В поиске указанной темы путеводителя использовались все перечисленные выше синонимы;
- нижестоящие (видовые) названия. Эти слова оказались полезными в двух ситуациях: когда у запроса имеется очень узкое или, напротив, излишне широкое содержание. При обработке узкого по содержанию запроса о *термическом старении бумаги* запрос был расширен при помощи термина *целлюлоза* (нижестоящего по отношению к слову *бумага*) и поисковых выражений, таких как *Тепловое старение целлюлозы, Термическое старение целлюлозы, Температурно-влажностное старение целлюлозы*. В рассмотренном поиске о бескислотной обработке бумаги (*Anoxic Treatment*) использовались такие ключевые слова, как *Ageless*^{TM} *, Oxygen Scavenger, Anoxic Cases, Anoxic Treatment Chamber, Anoxic treatment Systems, Low oxygen systems, Oxygen Absorber, Oxygen Scavenger, Veloxy System.* Обработку широкого по содержанию запроса с употреблением нижестоящих понятий иллюстрирует запрос *Биологические вредители*. Поиск по нему проводился с использованием нижестоящих понятий *Насекомые вредители книг, Насекомые вредители бумаги, Вредители в библиотеках, Вредители в музеях, Вредители в архивах, Крысы, мыши в библиотеках, Насекомые вредители складов, Насекомые в жилищах,*

Санитарный контроль помещений, Борьба с насекомыми в помещениях, Защита от насекомых в помещениях, Уничтожение насекомых в помещениях. Описанная обработка запроса дала высокую полноту результатов информационного поиска;

- другие условно эквивалентные термины. Эти термины связаны между собой в основном ассоциативным парадигматическим отношением, как *Книги – Библиотеки, Дезинфекция – Дезинфицирующие вещества, Фумигация – Фумиганты* и т.п.

Поиск по близким по содержанию словам может производиться при помощи операции дизъюнкции в логической форме сложных запросов или с использованием декомпозиции – разделения запроса на подзапросы. Различия этих двух операций описаны далее. Источником синонимии при подготовке запросов являлся тезаурус по сохранности документов [25], регулярные консультации со специалистами в области консервации документов и термины, получаемые из сетевой информации. Широкое употребление синонимов сделало актуальной задачу сохранения результатов описанной обработки. Запросы для комплектования путеводителя сохраняются в специальном файле запросов для их многократного употребления при проведении регулярного обновлении путеводителя.

5. Расширение с использованием наиболее употребительной формы запроса

Выбор самой употребительной формы запроса (*аспирин*, а не *ацетилсалициловая кислота*) имеет значение тогда, когда для желаемого результата нужен только один информационный запрос. Одного поиска и запроса обычно достаточно при получении справок различного бытового характера. В таких ситуациях выбор синонимов служит единственным средством обеспечения повышенной полноты: чем большую частоту имеет синоним, тем большая выдача в ответ на запрос. Другое дело – запросы, решающие не обиходные, а специальные поисковые задачи, как, например, работа по комплектованию информационных ресурсов и, в частности, путеводителя по сохранности документов. Эта работа ведется с использованием большого количества синонимов (иногда – десятков) без разделения их на редкие и часто используемые. Широким употреблением синонимов обеспечивается высокая полнота результатов информационного поиска и, соответственно, полнота информации в путеводителе. Различие в употребительности поисковых терминов становится важным тогда, когда поиск проводится для обновления путеводителя, и из возможных запросов отбираются для сохранения их наиболее употребительные формы из специального файла запросов. Малочастотные формы запросов при обновлении не используются, и таким образом уменьшается трудоемкость работы по обновлению путеводителя.

ДРУГИЕ ВОЗМОЖНЫЕ СРЕДСТВА РАСШИРЕНИЯ ЗАПРОСОВ

Следующие два правила составления запросов нами не рассматривались. Они сложились при наших собственных поисках в Интернете.

6. Расширение запроса с использованием его усечения. Усечение заключается в удалении части запроса, в результате чего получаются нулевые позиции (далее обозначаемые символом Ø) со значением ‘каждый’, ‘всякий’, ‘любой’ [26]. Это значение хорошо раскрывается при сравнении запросов “специализированные тезаурусы” и “тезаурус по определенной тематике”. Эти термины равноценны по смыслу. Название специализированный тезаурус означает тезаурус отраслевой или тезаурус по какой-либо области. Если форму запроса “тезаурус по определенной тематике” сократить до искусственной формы “тезаурус по [Ø]” (с усечением справа), то нулевая позиция в таком выражении эквивалентна дизъюнкции **всех конкретных названий** тезаурусов по их назначению. Эта декомпозиция включает такие названия, как *тезаурус по педагогике и образованию, тезаурус по этологии, социологии, рекурсии, транснациональному образованию, наукам о Земле, сельскому хозяйству* и многие, многие другие. Нулевая позиция имеет значение ‘любая предметная область’, а термины в приведенной дизъюнкции являются видовыми понятиями по отношению к родовому названию “специализированные тезаурусы”. Для поиска информации форма запроса, включающая нулевую позицию, имеет то преимущество, что она представляет собой наиболее “экономное” и простое языковое выражение запроса. При выборе этой формы запроса отпадает необходимость включения в запрос для его расширения большого количества (несколько сотен) конкретных названий тезаурусов. В силу того, что модель построения терминов “тезаурус по [область знания]” является наиболее распространенным обозначением понятия ‘специализированный тезаурус’, усеченная форма запроса также наиболее результативна.

Усечение запроса дает высокую полноту результатов информационного поиска. При создании путеводителя этот прием применялся при подготовке запроса по теме *Бумага ручного производства*, когда запрос был составлен, как *Бумага ручного* (в используемой нами нотации – *Бумага ручного* Ø). Ответом на этот вопрос была информация о бумаге *ручного изготовления, ручного литья, ручного отлива, ручного производства, ручной работы*. Нулевая позиция в приведенном запросе включает в себя содержание всех названных операций: *изготовление, литье, отлив, производство, работа*. Особенность этой позиции заключается в том, что Ø соответствует не только указанным выше пяти, но и любым операциям по ручному изготовлению бумаги.

7. Декомпозиция информационных запросов. Еще одно правило, широко применявшееся при создании путеводителя по сохранности документов, мы называем декомпозицией запросов. Декомпозиция как прием подготовки запросов путем их деления на подзапросы, ранее в настоящей статье не рассматривалась, но возможность ее вытекает из некоторых уже описанных правил. Согласно правилу 1 хорошие результаты при поиске дают узкие по содержанию запросы. И поскольку возможным путем получения узких запросов является разделение общих запросов на частные, декомпозицию можно рассматривать как особое правило составления запросов.

Декомпозиция близко стоит к рассмотренной как рекомендация к правилу 4, операции дизъюнкции поисковых терминов, называемой также отношением ИЛИ. Логическая форма с дизъюнкцией (например, *Anoxic treatment* OR *Anoxic storage* OR *Anoxic Treatment Chamber* OR *Anoxic Display* OR *Anoxic Cases* OR *Insect Control by Inert Gases* OR *Insect Control by Anoxia* OR *Anoxic Fumigation* OR *Low oxygen Treatment*) является полностью эквивалентной серии подзапросов, построенной с применением декомпозиции, как

Anoxic Treatment
Anoxic Storage
Anoxic Treatment Chamber
Anoxic Display
Anoxic Cases
Insect Control by Inert Gases
Insect Control by Anoxia
Anoxic Fumigation
Low Oxygen Treatment

Обе рассматриваемые формы соответствуют содержанию запроса “Бескислотная обработка”, в одинаковой степени служат для повышения полноты результатов информационного поиска и различаются в следующем.

При дизъюнкции терминов выдача информации по отдельным словам объединяется в одну общую выдачу, и из нее устраняются дубли – неоднократная выдача повторяющихся веб-страниц. В случае декомпозиции поиск проводится так же, но отсутствует объединение выдач и удаления дублей, в силу чего две последние операции выполняются самостоятельно, как специальная постобработка результатов информационного поиска. Второе, что различает рассматриваемые две операции – это сложности с применением дизъюнкции к излишне большому количеству поисковых терминов. Трудности возникают не при составлении запроса и проведении поиска, а при контроле полученной информации пользователем. При просмотре полученных веб-страниц, например, по запросу “Бескислотная обработка документохранилищ” – (*Anoxic treatment* OR *Anoxic storage* OR *Anoxic* OR *Insect Control by Anoxia* OR *Anoxic Fumigation* OR *Low oxygen Treatment*) *(Library OR Museum OR Archives OR Insect OR Pests)* – сопоставляются ключевые слова запроса со словами, имеющимися в получаемых документах. Такой контроль информации эффективен, когда контролируется небольшое количество терминов. По некоторым данным взрослые люди способны одновременно держать в уме от четырех до шести объектов, в данном случае – сопоставляемых слов [27, 28]. При сравнении большого количества поисковых признаков результаты поиска становятся с трудом контролируемыми просматривающим информацию пользователем.

При создании путеводителя по сохранности документов почти во всех случаях употреблялась декомпозиция, а дизъюнкция – редко и только в тех случаях, когда ею связывалось не более трех-четырех ключевых слов. Декомпозиция обеспечивает высокую полноту результатов информационного поиска, но в ситуациях, когда поиск ведется по 10 или нескольким десяткам запросов (а при создании путеводителя это вовсе не редкость), использование декомпозиции запросов отличается большой трудоемкостью.

В ходе работы с путеводителем декомпозиция имеет два назначения: создание классификационной схемы предметной области и подготовка запросов для проведения поиска. В классификационной схеме путеводителя (ниже приводится ее англоязычная часть), за исключением тематических классов *People & Organizations* и *Paper and Cardboard,* все остальные классы построены путем многократного разделения содержания понятия ‘сохранение документов’ на более частные рубрики (табл. 1).

Таблица 1

People & Organizations (10)	**Environment Management (730)**
Finding Conservation Organizations	Environment Control (250)
Finding People in Conservation	Indoor Air Quality, Dust (190)
Database of Conservators & Bookbinders	Lighting, UV & IR Radiation (170)
Paper and Cardboard (2320)	Temperature & Humidity (120)
Types of paper and paperboard (1520)	Pest Management (2820)
Paper & Cardboad Manufacture (340)	**Pest Control (310)**
Paper Acidity & Deacidification (210)	Integrated Pest Management (110)
Paper Aging (250)	Mold & Mildew (880)
Paper & Book Conservation (2650)	Mass Treatment (1110)
Reference Information (220)	Biocides/Pesticides (410)
General & Overview Information (140)	**Disasters & Incidents (950)**
Professional Standards (190)	Disaster Prevention (230)
Preventive Conservation (200)	Disaster Mitigation & Preparedness (200)
Bookbinding (500)	Disaster Response & Recovery (100)
Book Repair & Restoration (320)	Fire Protection (80)
Manuscripts Conservation (410)	Floods (80)
Leather Conservation (480)	Incidents, Theft, Vandalism (190)
Photographs Conservation (307)	Disaster resources (50)

Примечание: Цифры в скобках обозначают количество ссылок, стоящих за рубрикой.

Приведенные рубрики представляют собой верхний уровень иерархии понятий в рассматриваемой тематической области. Дальнейшие уровни получались при разделении вышеописанных рубрик на более мелкие классы. С использованием терминов тезаурусостроения можно сказать, что декомпозиция рубрик основана на использовании родо-видовых или иерархических отношений между понятиями. Декомпозиции содержания запросов основываются на отношении синонимии, поисковой эквивалентности терминов. Последний вид декомпозиции представлен следующими разновидностями.

(1) Декомпозиция, проводимая по синонимам и по названиям регулярных тематических классов. Регулярные классы запросов – это пять классов имеющейся в путеводителе информации, как *Bibliography, Dictionaries, Guidelines & Manuals, Databases* и *Directories*, отражающих жанровое разделение информации в путеводителе. Для термина *Mold* и его синонимов *Mould, Fungi* и *Mildew* этот уровень декомпозиции дает следующую серию, состоящую из 44 подзапросов:

Bibliography – Mold Bibliography, Mould Bibliography, Fungi Bibliography, Mildew Bibliography

Dictionaries – Mold Dictionary, Mould Dictionary, Fungi Dictionary, Mildew Dictionary; Mold Glossary, Mould Glossary, Fungi Glossary, Mildew Glossary; Mold Vocabulary, Mould Vocabulary, Fungi Vocabulary, Mildew Vocabulary

Guidelines & Manuals – Mildew Guidelines, Mould Guideline, Fungi Guideline, Mildew Guideline; Mold Guide, Mould Guide, Fungi Guide, Mildew Guide; Mold Manual, Mould Manual, Fungi Manual, Mildew Manual; Mold Tutorial, Mould Tutorial, Fungi Tutorial, Mildew Tutorial

Databases – Mold Database, Mould Database, Fungi Database, Mildew Database; Mold "Data base", Mould "Data base", Fungi "Data base", Mildew "Data base"

Directories – Mold Directory, Mould Directory, Fungi Directory, Mildew Directory.

(2) Декомпозиция по синонимам и лексическим привязкам запросов к тематической области консервации документов. Для указанного уточнения запроса, используются ключевые слова *Paper, Books, Documents, Libraries, Museums, Archives*. Для запроса "Плесень в хранилищах документов" результатом рассматриваемой декомпозиции является серия, состоящая из 24 подзапросов:

Fungi – Fungi Archives, Fungi Library, Fungi Museum, Document Fungi, Book Fungi, Paper Fungi

Mildew – Mildew Archives, Mildew Library, Mildew Museum, Document Mildew ...

Mold – Mold Archives, Mold Library, Mold Museum, Document Mold …

Mould – Mould Archives, Mould Library, Mould Museum, Document Mould …

Декомпозиции запросов и тематических рубрик имеют разное назначение. Назначением декомпозиции запросов является повышение полноты результатов информационного поиска. Широкое употребление при подготовке запросов синонимов в некоторых случаях – видовых (нижестоящих) понятий дает очень высокую полноту проведения поиска, сопоставимую с результатами поиска информации с употреблением тезауруса. Назначение декомпозиции рубрик – принципиально иное, при их помощи минимизируется наполнение тематических рубрик и работа пользователей с путеводителем делается более удобной. Удобство для пользователей заключается в том, что декомпозиция сокращает число документов, стоящих за рубрикой, уменьшает время просмотра ресурса и трудоемкость работы с путеводителем.

ВЫВОДЫ

В заключение приведем сводку правил, используемых при составлении запросов, и описанных в настоящей статье. Правила сгруппированы по их назначению – ориентации на достижение полноты или точности результатов информационного поиска (табл. 2).

Декомпозиция рубрик, по нашему мнению, не относится к собственно составлению запросов, а целиком лежит в области техники построения информационных ресурсов, включающих тематические рубрики и рубрикаторы информации. Описанные в настоящей статье правила составления запросов, по оценке известного специалиста по Яндексу Андрея Плахова, образуют язык, который «никто не использует, кроме некоторых специалистов» [29]. Данное мнение представляется излишне категоричным, в Интернете эти правила применяются широко, но их использование, в частности, для комплектования информационных ресурсов, далеко не всегда представляется очевидным. Запросы, построенные с применением названных правил, имеют относительно малое употребление в сети, поскольку они ориентированы не на повседневное (и поэтому массовое) употребление, а на специальные области знания. Указанным назначением определяется относительная редкость «необиходных» информационных запросов и рассмотренных правил их составления.

Таблица 2

Повышение точности поиска	***Повышение полноты поиска***
1. Уточнение содержания запросов ключевыми словами	4. Расширение запросов при помощи оператора ИЛИ (дизъюнкция)
2. Уточнение содержания запросов при помощи операторов	5. Использование наиболее употребительной формы запроса
3. Использование подсказок при наборе запросов	6. Расширение запроса путем его усечения
	7. Разделение запросов на подзапросы (декомпозиция запросов)

СПИСОК ЛИТЕРАТУРЫ

1. Поиск в Интернете, Все студенту. – URL: http://everything-for-students.ru/?p=306
2. Как пользоваться поисковиками. – URL: http://www.ooortk.ru/google
3. Как правильно искать в Google. – URL: https://lifehacker.ru/2012/09/28/kak-pravilno-iskat-v-google/
4. 7 поисковых операторов Google. – URL: http://monetavinternete.ru/vazhnye-temy/raznoe/7-poiskovyx-operatorov-google-kotorye-pomogut-oblegchit-poisk/
5. Академия Google: Советы по расширенному поиску в Академии Google. – URL: https://www. google.ru/intl/ru/scholar/refinesearch.html
6. Как искать информацию в Яндекс, Google: операторы поиска. – URL: http://www.businessvinternet.ru/biznes_v_internete/instrumenty_on_line_biznesa/700-kak-iskat-informaciyu-v-yandeks-google-operatory.html
7. Как искать в Гугле (Google). Операторы запросов. – URL: http://ci-razvedka.ru/Google_ Search_1.html
8. Поисковые операторы Google. – URL: http://seo-in.ru/poiskovaya-optimizaciya/74-operator-google.html
9. Правильный поиск в Google. Язык запросов в Google. – URL: http://www.diacr.ru/zametki/20-kak-pravilno-iskat-v-google/kak-pravilno-iskat-v-google.htm
10. Таблица операторов Google. – URL: http://archive.ec/0dgH
11. Язык расширенного поиска в поисковой системе Google. – URL: http://bourabai.ru/dbt/ google/query.htm
12. Google – «живой поиск» в строю… – URL: http://weno.ru/?p=114
13. Как искать в Интернете – Компьютер для новичков. – URL: http://beginpc.ru/internet/kak-iskat-v-internete
14. Искусство понимать с полуслова. Расширение запроса. – URL: https://habrahabr.ru/ company/ yandex/blog/187404/
15. Тема 4. Правила составления запросов. – URL: http://www.pravo.vuzlib.net/book_z1832_page_6.html
16. Тильда как оператор поиска больше не действует. – URL: http://ru-google-os.blogspot.ru/2013/06/google - tilde-operator-no-longer-works.html
17. В Интернете найдется все. Как правильно задавать вопросы поисковой системе? – URL: http://www.aif.ru/society/web/24747.
18. Эффективная методика поиска в Интернете с применением поисковых машин. – URL: http:// is-9.ru/laboratornaya-rabota-2-informatcionnye-seti
19. Ответы Mail.Ru: правила поиска информации в интернете! – URL: https://otvet.mail.ru/question/84586247
20. How googlit. – URL: http://ilch.vsmu.edu.ua/files/howgoog.htm
21. Расширенный поиск в Google. – URL: http://s3design.com.ua/rasshirennyj-poisk-v-google/
22. Правила и советы по поиску информации в Интернете. – URL: http://itandlife.ru/technology/poisk-informacii/
23. Как правильно искать в Google. – URL: http://today.vodafone.ua/posts/kak-pravilno-iskat-v-google
24. Путеводитель по вопросам сохранности документов на бумажной основе. – URL: http:// 91. 151.182.200: 8083/
25. Тезаурус информационно-поисковый по сохранности документов / сост.: Е.Н. Пименов, Л.Г. Левашова, В.Б. Никитин; Б-ка РАН. – СПб. – URL: http://www.rasl.ru/e_resours/tezaurus/index.htm
26. Пименов Е.Н. Предметно-аспектный подход к индексированию информации: актуальный и нулевой предмет индексирования // Научно-техническая информация. Сер. 2. – 2001. – № 7. – С. 18-25.
27. Википедия: Внимание. – URL: https://ru.wikipedia.org/wiki/%D0%92%D0%BD%D0%B8%D0%BC%D0%B0%D0%BD%D0%B8%D0%B5
28. Внимание человека: определение в психологии и характеристика. – URL: http://pamyatplus.ru/harakteristika/vnimanie-cheloveka.html
29. Язык поисковых запросов как естественный язык. – URL: https://events.yandex.ru/lib/talks/809/

Материал поступил в редакцию 07.11.16.

Сведения об авторе

ПИМЕНОВ Евгений Николаевич – кандидат филологических наук, старший научный сотрудник Отдела информатики и автоматизации Библиотеки Российской академии наук, Санкт-Петербург
e-mail: pen48@list.ru