

07

02

Информационно-поисковый тезаурус ИНИОН РАН по науковедению: наука и образование

О.М. КОРЧАЖКИНА

X

Востребованность информации в области науки, образования и новых технологий и необходимость обмена опытом между учеными и специалистами по науковедению подтверждается многолетними исследованиями, которые проводит Центр научно-информационных исследований по науке, образованию и технологиям Института научной информации по общественным наукам Российской академии наук (ИНИОН РАН).

С 1984 г. в ИНИОН РАН создаётся и постоянно обновляется уникальная по объёму и охвату предметных областей библиографическая база данных по науковедению, которая содержит около 250 тыс. документов – аннотированных указаний на монографии, сборники статей, авторефераты диссертаций, отдельные статьи из сборников, журналов и прочих периодических изданий на русском, западноевропейских, славянских и восточных языках. Последнее пополнение ресурса, датированное 1 февраля 2012 г., составило 9428 новых документов.

В условиях современного “информационного бума”, когда объём информационных потоков по всем отраслям знаний нарастает в геометрической прогрессии, особенно актуальной становится задача организации рациональной и эффективной системы поиска требуемой информации в текстовых базах данных. Для этого специалисты создают дополнительные ресурсы-посредники, называемые тезаурусами, которые обеспечивают возможность автоматизированного поиска требуемых ресурсов по вводимому запросу.

Информационно-поисковый тезаурус (ИПТ) – одна из разновидностей лингвистических ресурсов, используемых в системах информационного поиска на основе баз библиографических данных, нормализованных ключевых слов, рефератов или концептов, представляющих собой термины и понятия в определённой предметной области. Согласно действующим стандартам, ИПТ “это нормативный словарь, явно указывающий отношения между терминами и предназначенный для описания содержания документов и поисковых запросов” [Лукашевич 2011, 23].

При составлении тезаурусов как инструментов информационного поиска эти термины и понятия извлекаются из соответствующих текстов “вручную” путём так называемого ручного индексирования. Эта процедура достаточно трудоёмкая, что обусловлено наличием многоуровневых отношений между морфологически связанными лексемами, представленными в текстах, которые очень трудно поддаются формализации по ряду причин. Среди них специалисты отмечают следующие: обширный объём при содержатель-

ной неполноте ресурсов; трудности выявления сложных иерархических отношений между семантически связанными лексемами и, как следствие, трудности снятия омонимии, то есть многозначности слов; снижение качества информационного поиска при расширенных запросах, трудности отбора текстов, подвергаемых индексированию и, наконец, уровень квалификации и качество работы индексаторов [Там же, 10, 12].

Специалисты по разработке тезаурусов подчёркивают, что оптимальное число терминов как основных элементов ИПТ не должно превышать 10 тыс., причём дескрипторов¹ среди них должно быть не более 60–70%, а остаток покрывается аскрипторами (недескрипторами, т.е. синонимами или близкими по смыслу терминами). Это количественное соотношение обусловлено тем фактом, что, “с одной стороны, набор дескрипторов тезауруса должен быть достаточен для описания произвольного документа предметной области, с другой стороны, дескрипторов не должно быть слишком много, поскольку слишком большая величина тезауруса повышает субъективность индексирования и затрудняет развитие и использование тезауруса” [Там же, 38]. Н.В. Лукашевич приводит примеры более обширных тезаурусов, в которых учитывается специфика определённых предметных областей: “Тезаурус по архитектуре и искусству” содержит более 30 тыс. дескрипторов, что объясняется необходимостью индексирования не только информационных источников, но и музейных экспонатов; “Медицинский тезаурус MeSH” охватывает более 25 тыс. дескрипторов, что обусловлено необходимостью включения в ресурс “внутренней” терминологии многочисленных отраслей медицинского знания.

С целью унификации процедуры индексирования с 1970-х годов создаются национальные и международные стандарты разработки ИПТ, в которых определены области их применения, в частности, содержатся рекомендации по ограничению использования тезаурусов в автоматических системах индексирования [Там же, 12, 23]. Поэтому потребности обеспечения систем автоматизированного поиска соответствующими ИПТ заставляют разработчиков искать новые способы создания тезаурусов, предназначенных именно для автоматического индексирования и отличающихся от традиционных тезаурусов, составленных в соответствии с национальными и международными стандартами [Добров и др. 2009, 112].

Специалисты указывают, что для целей автоматического индексирования, т.е. при автоматической обработке текста согласно описанию его содержания в виде дескрипторов, традиционные ИПТ должны быть дополнены формализованными сведениями, которые обычно использует индексатор для определения основной темы текста. Эта процедура требует расширения зоны охвата специфических терминов путём составления соответствующих указателей – дополнительных словарей в составе ИПТ, что на практике обеспечивает переход к запросам по понятиям более высокого уровня [Там же, 119]. Поэтому для повышения качества пользовательских запросов и релевантности результатов поиска ИПТ, разрабатываемые для целей автоматического индексирования, должны содержать описание различных навигационных типов концептуальных отношений между входящими в него терминами в дополнительных синонимических указателях, называемых тезаурусными расширениями [Маннинг и др. 2011, 200–201]².

Существуют и другие подходы к созданию ИПТ, предназначенных для автоматического индексирования, которые основаны на статистических методах обработки текстов, а также на сводах правил, описывающих совокупность синтаксических и морфосинтаксических трансформаций, причём есть примеры, когда таких правил бывает установлено до 40 тысяч (!) для одного тезауруса [Добров и др. 2009, 116; Лукашевич 2011, 50].

Образцом удачно составленного ИПТ, предназначенного для автоматизированного поиска информации и содержащего тезаурусные расширения, является “Информационно-поисковый тезаурус ИНИОН по науковедению: наука и образование” [Информационно-поисковый тезаурус 2011] (ИПТ “Науковедение”). ИПТ “Науковедение” создан в ИНИОН РАН как основной компонент лингвистического обеспечения автоматизированной библиографической базы данных ИНИОН РАН по науковедению. На базе ИПТ “Науковедение” создана автоматизированная поисковая система по науковедению (<http://83.149.253.12/scripts/Rweb.exe?DBNAME=scien&DCNFN=152133&SYSLANG=RU>).

ИПТ "Науковедение" предназначен для индексирования входного потока документов и запросов и обеспечивает возможность интеллектуального поиска информации в библиографической базе данных ИНИОН. По тематическому охвату тезаурус ориентирован на следующие области: теорию и методологию науковедения, историю и современное состояние науки, науку и общество, организацию науки, высшую школу. ИПТ "Науковедение" как лингвистический ресурс включает отраслевую терминологию, терминологию других наук, рассматриваемую в науковедческом аспекте, общенаучную лексику. Авторы ресурса подчёркивают специфичность лексики тезауруса, обеспечивающую эффективную обработку документов и пользовательских запросов.

Тезаурус составлен по типу расширения на базе трёхуровневого указателя: алфавитного лексико-семантического, систематического и пермутационного указателя (общий объём полиграфического варианта – 494 с., формат 60 × 84/8, 62 усл. печ. л.). Наличие систематического и пермутационного указателей отличает ИПТ "Науковедение" от традиционных ИПТ, предназначенных для неавтоматического индексирования.

Алфавитный лексико-семантический указатель, являющийся основным, содержит 6627*словарных статей (терминов), из которых 4054 являются дескрипторами, а 2573 термина – аскрипторами. В словарных статьях дескриптора представлены его связи с другими дескрипторами и аскрипторами по типу недифференцированных иерархических отношений, ассоциативных отношений и отношений синонимии. В словарных статьях аскрипторов делается отсылка к соответствующим дескрипторам. Авторы указывают, что Алфавитный лексико-семантический указатель "обеспечивает возможность адекватной обработки запроса, а при неудовлетворительных результатах поиска – изменение стратегии поиска, в том числе за счёт использования дескрипторов, входящих в словарную статью" [Информационно-поисковый тезаурус 2011, 8].

В *Систематическом указателе* ИПТ "Науковедение" дескрипторы распределены в алфавитном порядке по рубрикам Рубрикатора ИНИОН и представлены в следующих разделах: общая лексика; науковедение; развитие современной науки, будущее науки; наука и общество; научный труд, научное творчество; организация науки, политика в области науки; методика и техника исследовательской работы; экономика науки; научные кадры; международное сотрудничество в науке; библиографические и справочные издания. Систематический указатель тезауруса используется для поиска дескрипторов в тематической рубрике при формулировке запроса, а также с целью повышения его релевантности.

Пермутационный указатель выполняет вспомогательные функции по отношению к Алфавитному лексико-семантическому указателю, поскольку содержит дескрипторы и аскрипторы без словарных статей. В Пермутационном указателе по однословным терминам и ключевым словам формируются словарные гнезда, иллюстрирующие употребление соответствующих лексических единиц в контексте, что позволяет использовать данный указатель для поиска нужных дескрипторов по ключевым словам в словарном гнезде.

В Предисловии к ИПТ "Науковедение" авторы подробно описывают структуру тезауруса и приводят примеры работы с каждым типом указателей [Там же, 5–11].

Таким образом, ИПТ "Науковедение" построен так, чтобы обеспечить поиск словарных статей по заглавным дескрипторам (с помощью алфавитного лексико-семантического указателя), поиск дескрипторов в рубриках (с помощью систематического указателя) и поиск дескрипторов и ключевых слов в словарном гнезде (с помощью пермутационного указателя). Эти возможности, заложенные его составителями, позволяют использовать тезаурус в системах автоматической индексации.

Если же рассматривать ИПТ "Науковедение" не как прикладной инструмент для обеспечения автоматизированного поиска информации по науковедению в библиографической базе данных ИНИОН РАН, а как классический тезаурус, т.е. словарь-сборник сведений, понятий, определений и терминов в области знаний по науковедению (ИПТ по науковедению содержит свыше 6600 терминов), то в доброй трети своих науковедческих материалов база данных может предложить философам самую новую информацию, базирующуюся на современных исследованиях и разработках, связанных с осмыслением структуры научно-

го знания, теорией познания, методологическими проблемами науки, историей науковедения. Термины “философия” и “философские науки” встречаются в базе данных в каждой 30-й работе, а если добавить термин “философские проблемы”, то в каждой 15-й работе.

Основное содержание рубрики “Теория и методология науковедения” более чем на 70% – это теория познания, логика и методология науки, философия науки. В рубрике “История науковедения. Персоналия” представлены как российские учёные – В.И. Вернадский, Б.М. Кедров, С.Р. Микулинский, М.К. Петров, И.Т. Фролов, В.С. Стёпин и другие, так и западные – Дж. Бернал, П. Бурдьё, М. Палани, Р.К. Мертон, Б. Рассел, Т. Кун, Ж.Ж. Соломон, Д. Прайс и другие. В рубрике “Теория науки. Методологические проблемы истории науки” содержатся материалы по философии естественных наук в целом и философии отдельных наук и научных дисциплин. История развития философии, философских наук, философской мысли составляет самостоятельный массив публикаций. Разделы “Будущее науки”, “Наука и общество”, “Современная научно-техническая революция”, “Наука и идеология”, “Наука и политика”, “Технократия”, “Глобальные проблемы человечества”, “Наука и нравственность”, “Наука и религия”, “Феномен интеллигенции”, “Социальная ответственность учёных” обширны по объёму единиц информации. Проблематика научного творчества рассматривается не только с позиций психологии индивидуальности, но и под философско-мировоззренческим углом зрения. В рубриках “Личность учёного”, “Проблемы общения в науке” размещены указания на публикации о философях различных стран мира, их научных трудах, научных школах и направлениях. В базе данных присутствуют материалы о современном состоянии исследований в области философских наук, о деятельности отдельных философских институтов, вузов, научных обществ. Отдельного внимания заслуживает информация о подготовке кадров философов, истории философских факультетов, деятельности аспирантуры и докторантуры. Самостоятельным массивом располагаются материалы о преподавании философии в вузах и школах, а также о международном сотрудничестве в области философии и философских наук.

ЛИТЕРАТУРА

Добров и др. 2009 – Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьёв В.Д. Онтологии и тезаурусы: модели, инструменты, приложения: Учебное пособие. М., 2009.

Информационно-поисковый тезаурус 2011 – Информационно-поисковый тезаурус ИНИОН по науковедению: наука и образование / РАН ИНИОН. Центр информатизации, Фундаментальная библиотека; Магай Е.В., Мдивани Р.Р., Хадияров Г.Г.; Ред.: Макешин Н.И., Мдивани Р.Р. М.: ИНИОН РАН, 2011.

Лукашевич 2011 – Лукашевич Н.В. Тезаурусы в системах информационного поиска. М., 2011.

Маннинг и др. 2011 – Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск / Пер. с англ. М., 2011.

Примечания

¹ Дескриптор – лексическая единица, авторизованный термин, входящий в ИПТ, т.е. однозначно поставленный в соответствие группе ключевых слов естественного языка, отобранных из текста, относящегося к определенной области знаний; словарная статья.

² Об эффективности поисковой системы, построенной на базе указанного тезауруса, можно судить по доле релевантных результатов поиска в общем наборе бинарных соотношений “запрос – документ” [Маннинг и др. 2011, 166], которая составляет 90–95%.