УДК 63:025.4

СИСТЕМА СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ МЕЖДУ ТЕРМИНАМИ ИНФОРМАЦИОННО-ПОИСКОВОГО ТЕЗАУРУСА ПО СЕЛЬСКОМУ ХОЗЯЙСТВУ И ПРОДОВОЛЬСТВИЮ

SYSTEM OF SEMANTIC RELATIONS BETWEEN TERMS OF INFORMATION RETRIEVAL THESAURUS ON AGRICULTURE AND FOOD INDUSTRY

Л. Т. ХАРЧЕНКО, кандидат сельскохозяйственных наук, старший научный сотрудник отдела аналитико-синтетической обработки документов
Л. Н. ПИРУМОВА, кандидат педагогических наук, заместитель директора, заслуженный работник культуры РФ
А. К. КОСТИН, кандидат сельскохозяйственных наук, старший научный сотрудник отдела аналитико-синтетической обработки документов
Ж. В. СОКОЛОВА, заведующая сектором
ФГБНУ «Центральная научная сельскохозяйственная библиотека» (ФГБНУ ЦНСХБ)

L. T. KHARCHENKO, candidate of agricultural science, senior scientist of department of analytical-synthetic documents processing
L. N. PIRUMOVA, candidate of pedagogical science, assistant director, honoured worker of culture RF
A. K. KOSTIN, candidate of agricultural science, senior scientist of department of analytical-synthetic documents processing
ZH. V. SOKOLOVA, chief of sector
FGBNU «Central scientific agricultural library» (FGBNU CSASL)

Описаны особенности тезауруса, создаваемого в ФГБНУ ЦНСХБ, и его характеристика. Рассмотрены методика работы с тезаурусом: отбор лексики, формирование словарной статьи, обработки терминов, работа с синонимией, принципы установления парадигматических связей термина, выстраивание иерархических деревьев.

Ключевые слова: лингвистическое обеспечение, базы данных, информационно-поисковые языки, тезаурусы.

Special aspects of the thesaurus created in the FSBSI CSAL and its characteristics are described. The following working techniques with the thesaurus were analyzed: selection of lexicon, formation of lexical entry, processing of terms, work with synonymy, principles of a term's paradigmatic connections establishment, formation of hierarchical trees.

Key words: linguistic support, databases, information retrieval languages, thesauruses.

Информационно-поисковые тезаурусы — самый сложный вид систематизированных терминологий. Это обусловлено установлением между его терминами нескольких видов семантических отношений в отличие от более простых терминологий (семантические ряды, таксономии), в которых между терминами устанавливается только один вид отношений — синонимичные или иерархические. Именно сложные семантические отношения, структурирующие понятия и термины предметной области, позволяют рассматривать тезаурус как модель науки, отрасли или области человеческой деятельности.

Особенностью Информационно-поискового тезауруса по сельскому хозяйству и продовольствию (ИПТ, или тезаурус) является широта предметной области, охватывающая естественные науки, возникшие на их основе прикладные науки, проблемы разных отраслей экономики, а также отраслей промышленности, связанных с сельскохозяйственным производством. ИПТ относится к информационно-поисковым языкам дескрипторного типа и используется при индексировании документов широкого тематического диапазона, поступающих на ввод в базу данных «АГРОС», генерируемую Федеральным государственным бюджетным научным учреждением «Центральная научная сельскохозяйственная библиотека». Объем БД «АГРОС» — более 1,8 млн записей на отечественные и зарубежные документы. ИПТ в настоящее время включает более 44 тыс. терминов. Из них 13 тыс. наименований организмов — растений, животных (в том числе рыб), микроорганизмов, являющихся объектами рассмотрения биологических наук, сельского, лесного, водного хозяйства и других областей.

Большая часть его лексического состава относится к сфере естественных и прикладных наук. Разработка и ведение ИПТ осуществляется на основе анализа текстового корпуса; понятия и термины рассматриваются с учетом их актуальности, частоты использования, представления в авторитетных лексикографических и электронных источниках, использования в документах электронных БД [3].

При создании больших тезаурусов, включающих понятия естественных наук, необходимо учитывать возникающие новые теории, пересмотр и замену устаревших понятий, что требует редактирования ИПТ, как с точки зрения семантики, так и с точки зрения корректности установленных семантических связей. Между терминами тезауруса установлено около 60 тыс. семантических связей трех видов: иерархические (отношения подчинения), синонимичные (отношения эквивалентности) и ассоциативные. Из них наиболее развиты иерархические отношения (около 30 тыс.).

Задействованные отношения важны не только с точки зрения контроля структуры тезауруса как терминосистемы, но и с точки зрения обеспечения с их помощью определенных удобств для быстрого нахождения в тезаурусе нужных терминов, автоматизированного

расширения поискового образа документа (ПОД) за счет подключения синонимичных рядов и иерархически вышестоящих терминов к дескрипторам, использованным в ПОД, что обеспечивает полноту информационного поиска.

Сравнительно просто в тезаурусе создаются синонимичные ряды. В специальной литературе и справочниках для некоторых понятий или объектов можно найти их упорядоченное представление с точки зрения синонимических связей и использовать его в ИПТ. Для других понятий и объектов источником синонимии служат индексируемые документы. Увеличение синонимичных рядов происходит по мере роста охвата информации определенной тематической направленности. В любом случае, используются сведения о синонимии, закрепленные в авторитетных источниках или научных текстах.

Иначе обстоят дела с отношениями подчинения (иерархии). Особенно — с ассоциативными отношениями, которые разрабатываются для лексических единиц конкретного тезауруса. В ИПТ использованы, в основном, два вида иерархических отношений: род-вид и часть-целое. Считается, что иерархические отношения в тезаурусах должны устанавливаться только в случаях, когда эти отношения истинны независимо от контекста. Например, согласно ботанической классификации, род Avena (овес) входит в сем. Poaceae (мятликовые, злаки), и эта связь существует независимо от контекста.

Следует отметить, что для отображения полезных и вредных организмов при индексировании документов БД «АГРОС» используется научный язык классификаций организмов (латинские названия), а также общеупотребительная лексика в виде русскоязычных терминов.

ИПТ в настоящее время содержит в качестве дескрипторов оба наименования сельскохозяйственных культур, дикорастущих растений, диких животных, промысловых и аквариумных рыб, что связано с необходимостью учитывать употребление разной лексики в биологических текстах и текстах хозяйственного плана, а также привязанность пользователя к традициям отечественной литературы.

В международных тезаурусах AGROVOC и CABI [1, 2] названия организмов представлены, за редким исключением, только научными терминами (латинские названия). Общеупотребительные названия, например, зерновых культур — пшеница, овес, ячмень при индексировании не используются, что в какой-то мере упрощает разработку и ведение тезаурусов, а также методику индексирования.

Установление иерархических отношений между терминами ИПТ — построение тех или иных иерархических деревьев обусловливается задачами индексирования и информационного поиска.

Так, отдельные виды рода Avena, например, овсюг (Avena fatua) являются растениями, которые засоряют посевы сельскохозяйственных культур и с ними ведется борьба. Другие виды, наоборот, служат объектом культивирования в качестве зерновой культуры как, например, овес посевной (Avena sativa).

Поиск по терминам сорняки и зерновые культуры актуален для БД «АГРОС» как информационного ресурса по вопросам сельского хозяйства. Учитывая это, а также удобство индексирования и расширения запроса по иерархии связей при поиске, созданы соответствующие иерархические деревья. При индексировании используется родовое имя (русскоязычный термин) и название конкретного вида растения (научное название), засоряющего посевы или культивируемого как сельскохозяйственная культура.

Иерархические отношения типа часть-целое в ИПТ, в основном, установлены между физическими объектами, процессами, свойствами, в сфере географического или административного деления. В большинстве случаев они представляют собой достаточно простые иерархические цепочки. Однако если понятие относится к области биологии, иерархические связи имеют тенденцию к усложнению. Так, сердце иерархически подчиняется двум вышестоящим терминам — кровеносная система и сердечно-сосудистая система, которые входят в иерархическое дерево анатомия животных.

Часто научно оправданное вхождение термина в несколько иерархических деревьев (полииерархия), с точки зрения задач поиска, нецелесообразно приводит к увеличению объема тезауруса. К тому же обилие автоматически приписываемых терминов по цепочке иерархии усложняет восприятие визуализированных ПОД. В связи с этим в ИПТ, как и в упомянутых выше международных тезаурусах, практикуется замена иерархических отношений ассоциативной связью. Ассоциативные отношения — наиболее сложные для определения. Считается, что основанием для их установления служит мысленное представление о существовании между двумя понятиями (терминами) какой-либо связи. Например, с точки зрения происхождения, целевого назначения, отношения к какому-нибудь процессу.

Отношения ассоциации устанавливаются только между дескрипторами, но никогда не связывают термины, имеющие между собой иерархические отношения или отношения синонимии. По принципу установления отношений ассоциативная связь симметрична. За ней закреплено также свойство наследования (транзитивности): она распространяется на все нижестоящие термины ассоциативно связанных дескрипторов.

В ИПТ в настоящее время контролируется только симметричность закрепления ассоциативной связи. Основными типами отношений, устанавливаемых между терминами в качестве ассоциативных, являются причина — следствие (загрязнение воздуха — парниковый эффект), болезнь — возбудитель (колибактериоз — Escherichia coli), сырье — продукт (табачное сырье — табак (продукция), процесс — назначение (обработка почвы — уход за растениями), процесс — объект (борьба с вредителями — вредители растений), объект — применение (почвообрабатывающие машины — обработка почвы), объект — свойства (поваренная соль — консерванты).

Ассоциативные отношения могут связывать дескрипторы, входящие в одну иерархию, дополняя друг друга каким-либо признаком, или частично совпадать по значению. В классификациях растений такая связь установлена между некоторыми родовыми именами, что отражает факт их ботанической близости, известный науке, или обладание некоторыми сходными признаками.

Ассоциативную связь в ИПТ используют в тех случаях, когда какой-либо род или вид дикорастущего растения еще только изучается с точки зрения декоративных, лекарственных, ядовитых или каких-либо других свойств для связи данного вида с определенной группой расте-

ний. После накопления достаточной информации о принадлежности данного рода или вида растений к хозяйственной группе принимают решение о замене ассоциативной связи иерархической.

Известно, что в тезаурусах различных предметных областей типы ассоциативных отношений будут разными, они могут не совпадать и в близких по тематике тезаурусах. В некоторых тезаурусах и лингвистических онтологиях традиционно используемую ассоциативную связь заменяют более четко обозначенными типами отношений.

С нашей точки зрения, традиционные ассоциативные связи между терминами, несомненно, важны и полезны — они отвечают задаче нахождения в ИПТ дополнительных терминов, полно и точно отображающих содержание документа или запроса. Кроме того, нередко они необходимы как элемент более точного определения другого термина.

● ЛИТЕРАТУРА

1. AGROVOC [Электронный ресурс]. — 2014. — Режим доступа: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus. — Загл. с экрана. *2.* CABI [Электронный ресурс], 2014. — Режим доступа: http://www.cabi.org/cabthesaurus/. — Загл. с экрана. *3. Пирумова Л. Н., Харченко Л. Т.* Тезаурус по сельскому хозяйству и продовольствию: индексирование документов и поиск информации в БД АГРОС: метод. Материалы. — М., 2001. — 68 с. 4.

e-mail: hlt@cnshb.ru, pln@cnshb.ru, sis@cnshb.ru