

ПРИНЦИПЫ РАЗРАБОТКИ ТЕЗАУРУСНОЙ БАЗЫ ДАННЫХ СЕЛЬСКОХОЗЯЙСТВЕННОЙ ТЕРМИНОЛОГИИ

Аширматова Манзура Жуманбаевна,

доцент

manzuraasirmatova274@gmail.com

Ташкентского государственного аграрного университета

Аннотация: В данной научной статье раскрываются принципы создания тезаурусной базы данных сельскохозяйственной терминологии. Подчеркивается, что создание тезаурусной базы данных является одним из наиболее сложных видов человеческой деятельности, требующим участия различных специалистов, включая лингвистов, специалистов в области информационных технологий и экспертов сельскохозяйственной отрасли.

Ключевые слова: *сельскохозяйственная терминология, создание тезаурусной базы данных, основные методы, алфавитный принцип, иерархический принцип, тезаурусное индексирование, ресурсные разделы.*

Abstract: This scientific article reveals the principles for creating a thesaurus database of agricultural terms. The development of a thesaurus database is considered one of the most complex types of human activity, requiring the participation of various specialists, including linguists, information technology experts, and agricultural professionals.

Keywords: *agricultural terms, creation of a thesaurus database, main methods, alphabetical principle, hierarchical principle, thesaurus indexing, resource sections.*

В настоящее время среди идеографических словарей особое внимание привлекают тезаурусы. Они находят своё применение не только в лингвистике, но и в сфере информационных технологий.

Тезаурус (греч. thesauros — «сокровище», «богатство») — это словарь, охватывающий всю лексику определённого языка и полно отражающий особенности её употребления в тексте. Изначально, в связи с постоянной

изменчивостью словарного состава живых языков, тезаурусы составлялись преимущественно на материале «мёртвых» языков либо на основе заранее определённых письменных памятников. В настоящее время тезаурусы представляют собой многогранные электронные лексикографические ресурсы, основанные на сложных терминологических и семантических отношениях, формирующих лингвистическую онтологию [1]. Тезаурусы являются лексикографическими источниками, охватывающими понятия, определения и термины определённой области знаний или деятельности, содержащиеся в текстовых корпусах или собраниях текстов.

Иными словами, тезаурус — это словарь, в котором слова и словосочетания с близким значением объединяются в единицы, называемые понятиями, концептами или дескрипторами, при этом между этими единицами отражаются их иерархические семантические отношения.

В тюркологии словари, составленные А.К. Боровковым, Е.И. Фозиловым, А. Наджибом на основе тюркских письменных памятников XIII–XIV веков, обладают признаками как глоссария, так и тезауруса. Словари, составленные по принципу отбора лексики в соответствии с языком отдельных авторов или их произведений, также относятся к тезаурусам. Кроме того, идеографическими словарями считаются такие, в которых лексические единицы определённой научной области расположены по тематическому принципу, а между ними отражены семантические отношения (родо-видовые, целое-часть, синонимические и др.). В таких словарях слова ищутся не по алфавиту, а по тематике и понятийному признаку.

С теоретической точки зрения тезаурус является одной из вероятностных моделей семантической системы. На практике он используется как средство обогащения индивидуального словаря автора. Кроме того,

тезаурус представляет собой специализированную базу данных, в которой формируются семантические сети.

Подобные словари служат инновационными лексикографическими ресурсами для создания онтологии естественного языка, обогащения лексикографических данных, проведения семантического анализа, машинного перевода, а также повышения эффективности образовательного процесса. Они являются важным инструментом в решении таких лингвистических проблем, как синонимия, омонимия, полисемия и определение фразеологических единиц.

Соотношение понятий лингвистической онтологии и тезауруса можно представить на следующей схеме (рис. 1).



Рис. 1. Графическое представление концептов лингвистической онтологии и тезауруса

Сегодня интеграция исследований и разработок в области NLP по мировым языкам повышает потребность в многоязычных тезаурусах. Существуют различные стандарты создания баз данных таких тезаурусов, а их

эффективность определяется наличием или отсутствием одноязычных тезаурусов. Для многоязычных тезаурусов, как правило, создаётся база данных, в которой для терминов из одного языка устанавливаются парадигматико-семантические отношения [рис. 2].

С этой точки зрения, в целях создания онтологии узбекского языка и многоязычных тезаурусов мы поставили задачу разработки тезауруса сельскохозяйственной терминологии на узбекском языке.

В русскоязычной научной практике в данной предметной области разработаны такие ресурсы, как Thesaurus CABI (<http://www.cabi.org/cabthesaurus/>) и AGROVOC (<http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>).

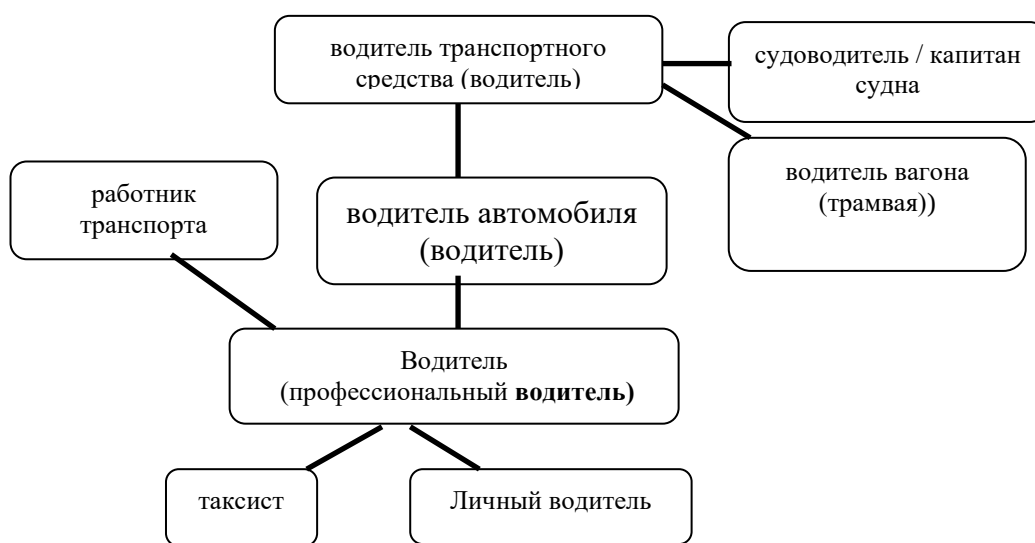


Рисунок 2. Концепты слова «водитель» в тезаурусе

Поскольку термины представляют собой общие лексические единицы для многих языков, данные системы могут быть использованы при создании тезаурусной базы данных сельскохозяйственной терминологии узбекского языка. Однако основная проблема данного подхода заключается в недостаточной разработанности терминологических переводных словарей.

Например, русско-узбекские источники сельскохозяйственной терминологии весьма ограничены, тогда как русско-английские и немецко-

английские ресурсы представлены значительно шире. Для обеспечения полноты и качества тезаурусной базы мы опираемся на существующие лексикографические источники (рис. 3).

акклиматизатор	ёмфирлаткич	оқпалак	тритикале
беркшир	закан	оқпоя	тухумқўсак
бетагабўз	зоотехния	оқселма	ўтлоқи
бешамаксиз	зооцидлар	оқшира	ўтлоқи-ботқоқи
бешбўлак	изасув	оқшўр	ферма
бонитировка	камшўр	пенициллиз	фумигантлар
гранозан	карпокинез	пестицидлар	фумигация
грена	квадрат-уялаб	пикировка	фунгицидлар
гренаж	кўзпайванд	пиретроидлар	фураж
ғон	кўп йиллик	плантаж	чигиттуширгич
ғўрасув	кўп пайкалли	пуркагичлар	чизель
дарактзорлаш	кўп тирсакли	раш	чиллаоб

Рис. 3. Пример ресурса для базы данных сельскохозяйственной терминологии. (<https://library.tdau.uz/uz/books/2174>)

Создание тезаурусной базы данных сельскохозяйственной терминологии включает следующие основные этапы:

- определение тематики тезауруса терминов;
- отбор словарей, отражающих сельское хозяйство, и создание базы данных тезауруса;
- формирование классификационной структуры понятий в области сельского хозяйства;
- установление семантических отношений между терминами (гипо-гиперонимия, меронимия, синонимия, омонимия и др.);
- пополнение толкового словаря сельскохозяйственных понятий через классификационную структуру и формирование на их основе тезауруса;

- формирование тезаурусной базы данных сельскохозяйственных терминов в алфавитном порядке и объединение её компонентов;
- совершенствование структуры публикации тезауруса и проведение экспериментального анализа с целью его расширения.

Отбор, определение и анализ терминов осуществляются с использованием методов классификации, сравнительного анализа, статистического анализа и компонентного анализа. Наиболее традиционным является эмпирический метод, при котором терминологические единицы выделяются на основе обработки текстовых документов по заданной тематике. Основными языковыми источниками служат технические и научные тексты, книги и публикации.

В ресурсной части базы данных количество взаимосвязанных терминов варьируется от языка к языку; эквивалентные термины в разных языках не всегда сопоставляются, и в каждом языке может быть представлен максимум один эквивалент. Таким образом, важно устранить несоответствия между многоязычными базами данных. Цель создания тезауруса заключается не в переводе терминов с одного языка на другой, а в установлении параметров, облегчающих информационный поиск в текстах.

Такой подход обеспечивает функционирование тезауруса как информационно-поискового тезауруса [1].

N.V. Lukashevich представила формальную модель информационно-поискового тезауруса следующим образом [Лукашевич, Добров, 2015]:

$$AQT = \langle D_{th}, T, R_H, R_A, A_T \rangle,$$

где:

D_{th} – множество дескрипторов предметной области, соответствующих понятиям конкретной предметной сферы; индекс “th” в данном случае означает, что авторами AQT в состав терминов-дескрипторов включены

единицы, необходимые для отражения основных тем ключевых документов, относящихся к данной предметной области;

T – множество терминов предметной области, при этом $D \subseteq T$;

RH – иерархические отношения информационно-поискового тезауруса;

RA – ассоциативные отношения информационно-поискового тезауруса;

AT – аксиомы транзитивности иерархических отношений.

При разработке тезауруса сельскохозяйственной терминологии были соблюдены следующие принципы: полнота охвата, формальный языковой аспект, согласованность и семантическая симметрия.

Полнота охвата: предполагается максимально полное включение терминов сельскохозяйственной отрасли. Для этого целесообразно опираться на лексикографические источники данной предметной области. Кроме того, для расширения охвата рекомендуется использовать тексты узбекских языковых корпусов, что позволяет включать термины, не отражённые в словарях и глоссариях. Это особенно характерно для нормативных документов.

Формальный языковой аспект: каждый термин, включённый в систему, относится к определённой части речи; теги частей речи чётко маркируются; осуществляется процесс лемматизации [Абжалова, 2020]. Это позволяет устранить языковые исключения и облегчает процессы машинного перевода и семантического анализа.

Согласованность: в соответствии с требованиями к тезаурусам обеспечивается согласованность определений абсолютных синонимов. Квазисинонимы и контекстуальные синонимы включаются в соответствующие синонимические синсеты только при их соотнесённости с конкретным абсолютным синсетом.

Семантическая симметрия: ключевой характеристикой тезауруса,

отличающей его от других лингвистических систем и лексикографических ресурсов, являются семантические отношения. Поэтому в тезаурусе сельскохозяйственной терминологии устанавливаются отношения род–вид, целое–часть, антонимия, омонимия и др., что обеспечивает системную связанность лексических единиц. В результате формируется развитая структура связей, повышающая эффективность информационного поиска.

В заключение следует отметить, что тезаурусы являются одним из наиболее эффективных инструментов описания предметных областей. В отличие от толковых словарей, тезаурус не только описывает значение слова, но и определяет его через систему связей с другими понятиями и их группами. Благодаря этому он эффективно используется для пополнения баз знаний систем искусственного интеллекта. В этой связи создание тезауруса сельскохозяйственной терминологии узбекского языка является актуальным как для разработки онтологии UzNet [Абжалова, 2022], так и для формирования многоязычных тезаурусов.

Литература

1. Abjalova M. O'zNet lingvistik ontologiyasini yaratish loyihasining konsepsiyasi. // O'zbekiston: til va madaniyat. Amaliy filologiya masalalari, – Toshkent: ToshDO'TAU, 1/2022. – B. 123-139.
2. Abjalova M.A. Axborot qidiruv tizimlarida tezauruslar ahamiyati. // Kompyuter lingvistikasi: muammolar, yechim va istiqbollar. Xalqaro ilmiy-amaliy konferensiya to'plami. Elektron nashr / ebook. – Toshkent: ToshDO'TAU, 22.04.2022. – B. 464-471.
3. Abjalova M.A. O'zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi. monografiya / M.A. Abjalova. –Toshkent: Nodirabegim, 2021. – 215 b. ISBN 978-9943-7804-5-3

4. Abjalova M. Tahriri va tahlil dasturlarining lingvistik modullari. [Matn] : monografiya / M.A.Abjalova. – Toshkent: Nodirabegim, 2020. – B. 25.
5. AGROVOC [Электронныйресурс]. – 2019. – URL: <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>.
6. [https://uz.wikipedia.org/wiki/Tezaurus#:~:text=Tezaurus%20\(yun.,holatlarini%20to%CA%BBlato%CA%BBkis%20aks%20ettiradigan%20lug%CA%BBat](https://uz.wikipedia.org/wiki/Tezaurus#:~:text=Tezaurus%20(yun.,holatlarini%20to%CA%BBlato%CA%BBkis%20aks%20ettiradigan%20lug%CA%BBat)
7. ThesaurusCABI [Электронный ресурс]. – 2018. – URL: <http://www.cabi.org/cabthesaurus/>.
8. THE TECHNOLOGY FOR CREATING A TERMINOLOGICAL THESAURUS AND PROSPECTS FOR ITS USE IN EDUCATION Ashirmatova Manzura Associate Professor of the Department of Uzbek Language and Literature, Tashkent State Agrarian University, UZ.
9. Ashirmatova Manzura Associate Professor of the Department of Uzbek Language and Literature Tashkent State Agrarian University, UZ LINGUISTIC SUPPORT OF THE THESEURUS OF AGRICULTURAL TERMS
10. Лавренова О. – Методика разработки информационно-поискового тезауруса. – М.: Пашков дом, 2001. – 54 с.
11. Лафtimi Имад. Информационно-поисковые тезаурусы: основные понятия, назначение и методика разработки. Отраслевой рыболовный тезаурус Текст: непосредственный // Молодой ученый. – 2012. –N 7 (42). – С. 164-166. – URL: <http://moluch.ru/archive/42/5096/> (дата обращения: 03.06.2020).