СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 006:81

А.Б. Антопольский

Международная стандартизация в сфере управления лингвистическими информационными ресурсами

Описываются задачи и результаты деятельности по стандартизации комитета ИСО/ТС37 «Язык и терминология», конкретно – задачи подкомитетов ПК3 и ПК4 по управлению терминологическими и другими языковыми ресурсами. Представлена классификация стандартов, раскрывается их содержание и приводятся аннотации. Основное внимание уделяется представлению лингвистической информации при разметке словарей и корпусов текстов, а также созданию лингвистических аннотаций к языковой информации.

Ключевые слова: *лингвистические ресурсы, стандартизация, терминологические базы данных, корпуса, лексиконы, языки разметки, лингвистические аннотации*

DOI: 10.36535/0548-0027-2021-05-5

ВВЕДЕНИЕ

Языковые технологии проникли в нашу повседневную жизнь. Для таких технологий необходимы разнообразные цифровые лингвистические информационные ресурсы (далее – ЛИР), а также стандартизированные процедуры их повторного использования и обмена.

При многоязычной коммуникации стандартизации требуют и огромные объемы производимой и используемой информации. Мировая языковая индустрия способствует многоязычному общению как в письменной, так и в устной форме.

Основная роль в стандартизации инструментов и ресурсов, используемых в языковой индустрии, принадлежит Техническому комитету 37 Международной организации по стандартизации «Язык и терминология» (ИСО/ТК37)[1]. Сфера деятельности этого комитета – стандартизация описаний, ресурсов, технологий и услуг, связанных с терминологией, письменным, устным переводом и другими языковыми видами деятельности в многоязычном информационном обществе.

В состав ТК37 входят 5 подкомитетов:

ПК1 Принципы и методы

ПК2 Рабочий процесс терминологии и языковое кодирование

ПК3 Управление терминологическими ресурсами

ПК4 Управление языковыми ресурсами

ПК5 Письменный, устный перевод и сопутствующие технологии.

В России действует аналог ТК37 – Технический комитет по стандартизации ТК55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле».

По состоянию на январь 2021 г. всего в ТК37 разработано и действует 70 стандартов и еще 31 находится в стадии разработки. Отмененные стандарты здесь не учитываются. Полный список стандартов ТК37 доступен по адресу https://www.iso.org/committee/48104/x/catalogue/p/0/u/1/w/0/d/0. В настоящем обзоре мы рассмотрим стандарты, разработанные подкомитетами ПК3 и ПК4, полностью посвященные управлению ЛИР.

Конечно, стандартами ТК37/ПК3 и ТК37/ПК4 сфера стандартизации лингвистических информационных ресурсов не исчерпывается. Существуют очень важные для ЛИР стандарты других комитетов ISO, прежде всего ТК46, а также других организаций, таких как IEEE, ISLE или W3C. Однако анализ всех этих документов выходит за рамки настоящей статьи.

В таблице приводится классификация семейств стандартов по управлению ЛИР.

1. Стандарты ИСО/ТК37/ПК 3 Управление терминологическими ресурсами

1.1. Системы управления терминологией, знаниями и содержанием

ISO 22274:2013[2] устанавливает основные принципы и требования для обеспечения применения классификационных систем во всем мире, учитывая

[1] ISO/TC37 – «Терминология и другие языковые ресурсы».

[2] ISO 22274:2013 Systems to manage terminology, knowledge and content – Concept-related aspects for developing and internationalizing classification systems. – URL: https://www.iso.org/ru/standard/36173.html

Таблица 1

Семейства стандартов по управлению ЛИР

ПК 3 Управление терминологическими ресурсами
- Системы управления терминологией, знаниями и содержанием
- Компьютерные приложения в терминологии
  - Структура терминологической разметки
  - Терминологические базы данных (в 3-х ч.)
- Реестр категорий данных (DCR)
- Обмен терминологическими базами (TBX)

ПК 4 Управление языковыми ресурсами
- Структуры функций
- Структуры компонентов
- Инфраструктура метаданных компонентов (CMDI) (в 2 ч.)
- Структура лексической разметки (LMF) (в 5 ч.)
- Пословная сегментация письменных текстов (в 2 ч.)
- Лингвистические аннотации
  - Структура лингвистических аннотаций (LAF)
  - Морфо-синтаксическая структура аннотаций (MAF)
  - Структура синтаксических аннотаций (SynAF) (в 2 ч.)
  - Структура семантических аннотаций (SemAF) (в 14-ти ч.)
  - Комплексная структура аннотаций (ComAF) (в 2-х ч.)
  - Справочная структура аннотаций (RAF)
- Многоязычная информационная структура (MILF)
- Постоянная идентификация и устойчивый доступ (PID)
- Контролируемый естественный язык (CNL) (в 4-х ч.)
- Инфраструктура компонентов метаданных (CMDI) (в 2-х ч.)
- Корпусный язык запросов LinguaFranca (CQLF) (в 2-х ч.)
- Транскрипция устной речи

такие аспекты, как культурное и языковое разнообразие, а также требования рынка. Стандарт учитывает, что системы классификации должны подходить для использования во всем мире и могут быть адаптированы к конкретным сообществам пользователей. В нем содержится информация о разработке и использовании классификационных систем, которые в полной мере подходят для различных языковых, культурных и рыночных условий.

Стандарт определяет факторы, которые необходимо учитывать при создании и заполнении системы классификации для использования в различных лингвистических средах. Эти факторы содержат спецификацию принципов включения аспектов интернационализации в классификационные системы, а также сохранение и использование этих аспектов для структурирования субъектов компании или организации.

В сферу действия стандарта ISO 22274:2013 входят: а) руководящие принципы по информационному содержанию для поддержки интернационализации систем классификации и лежащих в их основе концептуальных систем; б) терминологические принципы, применимые к системам классификации; в) требования к интернационализации систем классификации; г) соображения по документообороту и администрированию содержания систем классификации для использования во всем мире.

1.2. Компьютерные приложения в терминологии

1.2.1. Структура терминологической разметки (TMF)

Центральным для терминологических ресурсов является стандарт ISO 16642:2017[3], который устанавливает все формы терминологической разметки.

Этот стандарт определяет структуру, которая включает метамодель и методы для описания конкретных языков терминологической разметки (TML). Стандарт разработан для поддержки разработки и использования компьютерных приложений для терминологических данных и обмена такими данными между различными приложениями. В документе установлены условия, которые позволяют отображать данные, выраженные в одном TML, на другой TML, и для этой цели формирует универсальный инструмент отображения

Конкретный TML может быть описан набором характеристик:

- отражение структурной организации метамодели (т. е. деревьев расширения TML),
- конкретные категории данных, используемые TML, и их отношение к метамодели,

[3] ISO 16642:2017 Management of terminology resources. Terminological Markup Framework (TMF). – URL: https://www.iso.org/standard/56063.html

- способ, которым эти категории данных могут быть выражены в XML и, таким образом, привязаны к деревьям расширения TML;
- словари, используемые TML для выражения таких различных информационных объектов, как XML-элементы и атрибуты, согласно соответствующим стилям XML.

TMF существует в XML, UML и в RDF, что дает различные возможности по моделированию метаданных. Более подробное описание TMF можно найти также в работе [1].

1.2.2. Терминологические базы данных

Новый стандарт **ISO 26162:2019 «Управление терминологическими ресурсами – Терминологические базы данных»** (Management of terminology resources – Terminology databases), заменивший стандарт ISO 26162:2010, состоит из двух частей:

ISO 26162-1:2019 – Часть 1: Проектирование»[4].

ISO 26162-2:2019 – Часть 2: Программное обеспечение»[5].

В настоящее время в разработке находится 3-я часть этого стандарта – **ISO/CD 26162-3** – Часть 3: Содержание[6].

Терминологические данные собираются, организуются и хранятся в самых разнообразных системах управления терминологией (TMS), которые используют различные системы управления базами данных – от приложений в персональном компьютере до веб-приложений в крупных компаниях и правительственных учреждениях. Серия ISO 26162 содержит рекомендации по разработке терминологических баз данных и основных функций TMS.

Функции извлечения терминов и выявления новых терминов стали предметом разработки нового стандарта[7].

1.3. Реестр категорий данных (DCR)

ISO 12620:2019 и его российский аналог ГОСТ Р ИСО 12620-2012[8] содержат руководящие принципы и требования, регулирующие спецификации категорий данных для ЛИР. Он определяет механизмы создания, документирования, согласования и поддержания спецификаций категорий данных в репозитории категорий данных (DCR). Спецификации категорий данных описывают отдельные информационные блоки, определяющие схему сбора или аннотирования данных для конкретных ЛИР. Каждая спецификация задает формальное представление категории данных и включает конкретные признаки, описывающие эту категорию (например, ее имя, определение, примеры, комментарии и т.д.). Кроме того, спецификация формирует порядок ее создания и ведения в реестре DCR. Группы категорий данных, которые выделены в качестве подмножеств их глобального набора, составляющего реестр DCR, образуют выборки категорий данных (DCS), в которых наряду с моделью данных должны быть определены различные ограничения, применимые к информационным структурам или форматам обмена, специфическим для тематической области или приложения.

Проблемы, связанные с применением DCR, изложены в работе [2].

1.4. Обмен терминологическими данными (TBX)

Стандарт ISO 30042:2008[9] и его российский аналог ГОСТ Р ИСО 30042-2016 определяют структуру TBX, разработанную для поддержки разных типов терминологических данных, включая анализ, описательное представление, распространение и обмен в различных информационных средах. Основная цель TBX — обмен терминологическими данными. Области применения – перевод и создание терминологических баз данных.

В стандарте представлены: требования к TBX-файлам, модульная структура ядра записи, метаданные модуля основной структуры и др. Описан также TBX-Basic – более легкая версия TBX, особенно подходящая для малых предприятий и для языковых приложений, требующих упрощенного подхода к управлению терминологией. TBX-Basic-это совместимый с TBX язык разметки терминологии, который допускает ограниченный набор категорий данных.

2. Стандарты ИСО/ТК37/ПК4 Управление языковыми ресурсами

2.1. Общие сведения

Содержание данного раздела строится на основе публикаций [3-5], а также используются аннотации действующих стандартов ПК4.

Цель ИСО/ТК37/ПК4 – подготовка международных стандартов и руководящих принципов для эффективного управления ЛИР в многоязычном информационном обществе. Комитет разрабатывает принципы и методы создания, кодирования, обработки и управления ЛИР, такими как письменные и речевые корпуса, лексиконы, словари и классификации. Основное внимание уделяется моделированию данных, аннотированию, разметке, обмену данными и

---

[4] ISO 26162-1:2019 Management of terminology resources – Terminology databases – Part 1: Design. – URL: https://www.iso.org/standard/71941.html и https://www.iso.org/obp/ui/#!iso:std:71941:en .

[5] ISO 26162-2:2019 Management of terminology resources – Terminology databases – Part 2: Software). – URL: https://www.iso.org/standard/71942.html и https://www.iso.org/obp/ui/#!iso:std:71942:en

[6] ISO/CD 26162-3 Management of terminology resources – Terminology databases – Part 3: Content. – URL: https://www.iso.org/standard/80464.html

[7] ISO/AWI 5078 Management of terminology resources – Terminology extraction. – URL: https://www.iso.org/standard/81917.html

[8] Management of terminology resources – Data category specifications. – URL: https://www.iso.org/standard/69550.html ГОСТ Р ИСО 12620-2012 Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов. – URL: http://docs.cntd.ru/document/1200104401

[9] ISO 30042:2008 Management of terminology resources – TermBase eXchange (TBX). – URL: http://www.ttt.org/oscarStandards/tbx/tbx_oscar.pdf ; ГОСТ Р ИСО 30042-2016. Системы управления терминологией, базами знаний и контентом. Обмен терминологическими базами. – URL: http://docs.cntd.ru/document/1200142747. Предлагаемый раздел представляет собой сокращенное и отредактированное изложение фрагментов российского ГОСТа.

оценке ЛИР, отличных от терминологических ресурсов, которые разрабатываются в ИСО/ТК37/ПК3.

Известно, что многие сообщества скептически относятся к навязыванию каких-либо стандартов вообще и выдвигают два основных аргумента.

Во-первых, многообразие теоретических подходов, в частности, к обозначению различных языковых явлений, позволяет предположить, что стандартизация по меньшей мере непрактична.

Во-вторых, существует опасение, что огромные объемы существующих данных и программное обеспечение для их обработки, на создание которого, возможно, потребовались годы усилий и значительные финансовые средства, будут полностью уничтожены принятием новых стандартов.

Чтобы устранить эти опасения, члены ПК4 направляют усилия на выявление моделей и общих рамок для создания и представления ЛИР. Эти модели, в принципе, должны быть достаточно абстрактными, чтобы вместить различные теоретические подходы.

Создавая стандарты на основе XML, подкомитет стремится обеспечить их совместимость с устоявшимися и широко принятыми веб-технологиями. Модели, разработанные до сих пор, демонстрируют, что это не является нереализуемой целью и возможен переход от устаревших форматов к форматам XML.

В настоящее время языковые специалисты недостаточно информированы об усилиях по стандартизации, предпринимаемых ИСО/ТК37/ПК4. Повышение их осведомленности о результатах и возникающих проблемах будет решающим фактором успеха комитета, необходимым для обеспечения широкого применения разрабатываемых им стандартов. Цель ИСО/ТК37/ПК4 заключается в разработке платформы для проектирования и реализации форматов ЛИР с целью облегчения обмена информацией между приложениями NLP. Это будет достигнуто путем определения общего интерфейса, способного представлять различные виды лингвистической информации.

Формат интерфейса должен отвечать следующим требованиям:

- *выразительность*: формат должен быть достаточно богатым, чтобы представлять все разновидности лингвистической информации;
- *семантическая адекватность*: данные формата должны иметь формальную семантику, т. е. их определение должно обеспечивать строгую основу для дальнейшей обработки;
- *инкрементальность*: поддержка различных стадий интерпретации входных данных и генерации выходных данных, позволяющая осуществлять как раннее, так и позднее слияние и деление;
- *единообразие*: представление различных типов ввода и вывода данных должно использовать одни и те же “строительные блоки” и одни и те же методы для объединения сложных структур, состоящих из этих строительных блоков.

Структура интерфейса должна соответствовать развивающейся области проектирования языковых систем, удовлетворяя следующим дополнительным требованиям:

- *открытость*: структура должна обеспечивать возможность репрезентаций, основанных на различных теориях и подходах;
- *расширяемость*: формат должен быть совместим с альтернативными методами проектирования репрезентаций (например, XML).

Архитектура данных для представления лингвистической информации включает:

- *базовые компоненты*: базовые конструкции для построения репрезентаций лингвистической информации, в частности, определение типов строительных блоков и способов их соединения;
- *общие механизмы*: методы представления, которые делают аннотации более компактными и гибкими и позволяют связывать их с внешними источниками информации;
- *контекстуальные категории данных*: административные (мета-)данные, релевантные для обработки, такие как данные окружающей среды (например, временные метки, пространственная информация); обрабатываемая информация; интерактивная информация (спикер, аудитория и т. д.).

Работа в ПК4 строится на основе рабочих групп:

- Терминология ЛИР
- Лингвистические аннотации
- Метаданные для мультимодальной и многоязычной информации
- Представление структурного содержания (синтаксис и морфология)
- Мультимодальное представление контента
- Представление уровня дискурса
- Многоязычное текстовое представление
- Лексиконы
- Валидация языковых ресурсов
- Сетевая коллаборация создания ЛИР

2.2. Центральная роль руководящих принципов TEI [6]

Подкомитет возник вслед за руководящими принципами (правилами) TEI [7], ставшими одним из самых эффективных приложений XML в мире цифровых гуманитарных проектов, а также крупномасштабных документальных систем, таких как двести миллионов документов Европейского патентного ведомства. В совокупности правила TEI имеют две взаимодополняющие особенности, которые делают их важными участниками процесса стандартизации документов:

- XML-словарь из почти 600 элементов, который охватывает большинство потребностей, связанных с текстом, и большинство текстовых жанров, квалифицируемых как языковые ресурсы: проза, поэзия, драма, транскрипция рукописей и разговорный дискурс;
- правила TEI основаны на платформе спецификации ODD (One Document Does it All), что обеспечивает легкую настройку правил в соответствии с потребностями.

Правила TEI обеспечивают модульную структуру специфичных для ЛИР стандартов, главным образом, для корпусного представления, разметки и аннотации. TEI предлагает очень гибкие механизмы, которые на практике приводят к тому, что возникает большое разнообразие упрощенных подмножеств.

В целом, правила TEI сыграли центральную роль в разработке стандартов ИСО/ТК37/ПК4 на различных уровнях. Был опубликован совместный стандарт

на представление структуры признаков (ISO 24610-1 и -2). Некоторые стандарты, такие как MAF, по умолчанию предлагают реализацию с использованием словаря TEI. Механизмы TEI (такие как указатели и конкретные элементы) были вставлены для обеспечения локальной совместимости между различными форматами аннотаций. Язык спецификации использован для разработки новых словарей.

Правила TEI стали основным инструментом создания систем разметки для корпусной лингвистики. Подробное описание их применения представлено в работе В.П. Захарова [8]. Версия правил TEI, адаптированная для русского языка, легла в основу модели разметки Национального корпуса русского языка [9].

2.3. Структуры признаков

ISO 24610-1:2006[10] – устанавливает формат для представления, хранения и обмена структурными признаками в приложениях, связанных с аннотированием, производством или анализом лингвистических данных, а также для описания ограничений, которые имеют отношение к набору признаков, значениям признаков, спецификациям признаков и операциям над структурами признаков. Таким образом, создается средство проверки соответствия каждой структуры признакам эталонной спецификации.

ISO 24610-2:2011[11] – устанавливает формат для представления, хранения или обмена структурными объектами в приложениях на естественном языке, как для аннотирования, так и для производства лингвистических данных. Стандарт предоставляет компьютерный формат для определения иерархии типов и ограничений, которые накладываются на набор спецификаций объектов и операций над структурами объектов. Структуры признаков – это существенная часть многих лингвистических формализмов. Формат может быть использован для документирования любой системы признаков, но в первую очередь – для типового представления структуры признаков, определенного в ISO 24610-1.

2.4. Структура лексической разметки (LMF).

Семейство стандартов LMF первоначально началось со стандарта ISO 24613-2008[12], позже отмененного. Новая версия стандарта представлена в виде 5 частей:

ISO 24613-1: 2019[13] Ч.1.: Базовая модель

ISO 24613-2: 2020[14] Ч.2.: Модель машиночитаемых словарей (MRD)

ISO/FDIS 24613-3:2020[15] Ч.3 Этимологические расширения

ISO 24613-4[16]Ч.4.: Сериализация TEI

ISO/DIS 24613-5[17]Ч.5. Сериализация обмена лексическими базами

Цель LMF – предоставить общую модель для создания и использования лексических ЛИР, для управления обменом данными между ними, а также для объединения отдельных ЛИР в глобальные ЛИР.

Конкретные экземпляры стандартов LMF могут содержать одноязычные, двуязычные или многоязычные ЛИР. Одни и те же спецификации должны использоваться для малых и больших, простых и сложных лексиконов, письменных и устных лексических представлений. Их охват не ограничивается только европейскими языками, они охватывают все естественные языки. Средствами LMF можно представлять большинство существующих лексиконов.

Двухуровневая организация образует связную семью стандартов со следующими общими и простыми правилами:

- спецификация высокого уровня предоставляет структурные элементы, которые украшены стандартизованными константами;
- спецификации низкого уровня предоставляют стандартизованные константы в виде метаданных.

LMF содержит следующие компоненты:

- базовый пакет, который представляет собой структурный каркас, описывающий базовую иерархию информации в лексической статье;
- расширения основного пакета, которые выражаются в структуре, описывающей повторное использование основных компонентов в сочетании с дополнительными компонентами, необходимыми для конкретного лексического ресурса;
- расширения, специально предназначенные для морфологии, машиночитаемых словарей, различных уровней NLP, многословных выражений.

Первым шагом в разработке LMF было создание общей структуры, основанной на общих особенностях существующего инструмента сегментации текстов, и последовательной терминологии для описания компонентов этих лексиконов. Следующий шаг – это проектирование всеобъемлющей модели, которая наилучшим образом представляла бы особенности различных лексиконов.

Приведем для примера аннотации отдельных частей стандарта LMF.

ISO 24613-1:2019 – описывает базовую модель LMF, метамодели для представления данных в одно-

[10]ISO 24610-1:2006 Languageresourcemanagement – Featurestructures – Part 1: Feature structure. – URL: https://www.iso.org/ru/standard/37324.html

[11] ISO 24610-2:2011 Language resource management – Feature structures – Part 2: Feature system. – URL: https://www.iso.org/standard/43823.html

[12] ISO 24613-2008 Language resource management – Lexical Markup Framework (LMF). – URL: http://www.lexicalmarkupframework.org/; http://en.wikibedia.ru/wiki/

[13] ISO 24613-1: 2019 Language resource management – Lexical markup framework (LMF) – Part 1: Core model. – URL: https://www.iso.org/ru/standard/68516.html

[14] ISO 24613-2: 2020 Language resource management – Lexical markup framework (LMF) – Part 2: Machine-readable dictionary (MRD) model. – URL: https://www.iso.org/standard/75407.html

[15] ISO/FDIS 24613-3: 2020 Language resource management – Lexical markup framework (LMF) – Part 3. Etymological extension. – URL: https://www.iso.org/standard/75407.html

[16] ISO 24613-4 Language resource management – Lexical markup framework (LMF) – Part 4: TEI serialization. – URL: https://www.iso.org/obp/ui/#iso:std:iso:24613:-4:ed-1:v1:en

[17] ISO/DIS 24613-5 Language resource management – Lexical markup framework (LMF) – Part 5: Lexical base exchange (LBX) serialization. – URL: https://standards.iteh.ai/catalog/standards/iso/d3e92229-20a1-45ca-9b6b-6a8d3a339b95/iso-dis-24613-5

язычных и многоязычных лексических базах данных, а также механизмы, позволяющие разрабатывать и интегрировать различные типы ЛИР.

ISO 24613-2:2020 – устанавливает модель машиночитаемого словаря (MRD), метамодель для представления данных, хранящихся в электронных словарях, предназначенных как для прямой поддержки переводчика, так и для машинного перевода.

ISO 24613 – описывает сериализацию модели LMF, определенной как XML-модель, совместимую с руководящими принципами TEI. Сериализация представляет собой процесс преобразования объекта в форму, подготовленную для передачи. Например, можно сериализовать объект и передать его по Интернету с использованием протокола HTTP между клиентом и сервером. И, наоборот, при десериализации объект воссоздается из потока. При XML-сериализации в поток XML сериализуются только открытые поля и значения свойств объекта. Эта сериализация охватывает классы базовой модели LMF, а также классы, предоставляемые следующими дополнительными частями ISO 24613: машиночитаемые словари и этимологическое расширение.

ISO/DIS 24613-5 – описывает сериализацию модели LMF, определенной как XML-модель, производную от схемы LBX и совместимую с XML-схемой. Эта сериализация охватывает классы данных, аналогично предыдущему стандарту.

Подробное описание LMF представлено в отдельной книге [10].

2.5. Пословная сегментация письменных текстов

ISO 24614-1:2010[18] и российский аналог **ГОСТ Р ИСО 24614-1-2013** представляют основные понятия и общие принципы пословной сегментации.

В языковых технологиях слово – это фундаментальное и необходимое понятие. Поэтому для сегментации текста на слова важно иметь универсальное определение слова. Нельзя просто использовать для разграничения слов правила, основанные на идентификации пробелов и знаков пунктуации. Такие правила не учитывают случаи сложных слов, которые пишутся через дефис, сокращений, идиом или словоподобных выражений, содержащих символы или цифры.

Представим некоторые применения и сферы, которые требуют сегментировать тексты на слова и к которым, следовательно, применим данный стандарт.

Перевод. Подсчет слов – главный метод оценки стоимости перевода.

Управление контентом. Большинство систем и баз данных для управления информационным содержанием (контентом) предусматривает поиск по отдельным словам.

Технологии распознавания речи. Системы речевого воспроизведения текста синтезируют речь на базе слов и поэтому требуют пословной сегментации для обеспечения возможности словарного поиска, расстановки ударений, установки просодического образца и др.

Прикладная лингвистика. Различные системы обработки текстов на естественных языках (NLP) должны сегментировать текст на слова для того, чтобы выполнить свои функции.

Лексикография. Лексические ресурсы часто оцениваются по их объёму – обычно на основе подсчёта числа слов. Объём языковых ресурсов – весьма важный показатель для управления ими.

Данный стандарт имеет часть 2, посвященную сегментации китайского, японского и корейского языков[19].

2.6. Лингвистические аннотации

2.6.1. Общая модель лингвистической аннотации

Лингвистическая аннотация содержит любые описательные или аналитические сведения (примечания), применяемые к необработанным языковым данным.

Основные данные могут быть динамическими (аудио, видео) или текстовыми. Добавленные примечания могут включать в себя транскрипции всех видов (от фонетических характеристик до структур дискурса), теги частей речи и смысла, синтаксические данные, идентификацию «именованных объектов», аннотации со ссылками и так далее. Большое внимание уделяется инструментам и форматам, которые широко используются для создания аннотированных лингвистических баз данных.

Лингвистические аннотации – наиболее разработанное направление деятельности ИСО/ТК37/ПК4, которое включает 20 документов, объединенных в следующие стандарты:

- Структура лингвистических аннотаций (LAF)
- Морфо-синтаксическая структура аннотаций (MAF)
- Структура синтаксических аннотаций (SynAF) (в 2-х ч.)
- Структура семантических аннотаций (SemAF) (в 14-ти ч.)
- Комплексная структура аннотаций (ComAF) (в 2-х ч.)
- Справочная (референтная) структура аннотации (RAF).

Фундаментальное положение семейства стандартов для лингвистических аннотаций состоит в том, что репрезентативные формы для лингвистических данных и их аннотаций могут быть смоделированы путем объединения структурной метамодели, т. е. абстрактной структуры, общей для всех документов данного типа (например, синтаксической аннотации), с набором категорий данных, которые связаны с различными компонентами метамодели. Деятельность ПК4 направлена на идентификацию сокращенного набора метамоделей, которые могут быть использованы для любого типа лингвистических данных и их аннотаций.

[18] ISO 24614-1: 2010 Language resource management – Word segmentation of written texts – Part 1: Basic concepts and general principles. – URL: https://www.iso.org/ru/ standard/ 41665.html; ГОСТ Р ИСО 24614-1-20 Менеджмент языковых ресурсов. Пословная сегментация письменных текстов. Часть 1. Основные концепции и общие принципы. – URL: http://docs.cntd.ru/document/1200108539

[19] ISO 24614-2: 2011 Language resource management – Word segmentation of written texts – Part 2: Word segmentation for Chinese, Japanese and Korean. – URL: https://www.iso.org/ru/standard/41666.html

Категории данных, определяются разработчиком; взаимодействие между форматами обеспечивается с помощью Реестра категорий данных (см. п. 2.3), в котором точно определены категории и отношения, необходимые для конкретного типа обозначения.

Модель лингвистической аннотации должна удовлетворять двум основным критериям:

- иметь возможность создавать аннотацию с использованием стандартного формата представления;
- служить в качестве базового формата, из которого могут быть выведены любые новые форматы для сравнения и объединения данных, а также для работы с данными с помощью общих инструментов.

Приведем примеры аннотаций стандартов, относящихся к этому семейству.

ISO 24612:2012[20] – определяет структуру лингвистических аннотаций (LAF) для представления языковых данных, таких как корпусы текстов, записи речи и видео. Предлагаемая структура включает в себя абстрактную модель данных и XML-сериализацию этой модели для представления аннотаций первичных данных.

ISO 24611:2012[21] – обеспечивает структуру морфо-синтаксических аннотаций (MAF) лексем, их связей и морфо-синтаксических свойств. MAF опирается на категории данных, содержащиеся в Реестре категорий данных ISOCat (DCR по ISO 12620); в нем описывается XML-сериализация MAF, согласно правилам TEI.

ISO 24615-1:2014[22] и российский аналог ГОСТ Р ИСО 24615-2016 – основаны на эталонных моделях и форматах представления синтаксической информации, являющейся результатом работы синтаксического анализатора или аннотациями ЛИР в банках деревьев. Стандарт описывает структуру синтаксических аннотаций (SynAF), высокоуровневую модель представления синтаксических аннотаций лингвистических данных с целью поддержки взаимодействия между языковыми ресурсами или компонентами языковой обработки. Заметим, что банков деревьев, построенных по этой модели, достаточно много (см. например, их перечень в Википедии – https://en.wikipedia.org/wiki/Treebank).

ISO 24615-2:2018[23] – описывает XML-совместимую сериализацию метамодели ISO 24615-1 с целью поддержки взаимодействия между ЛИР или компонентами языковой обработки в области синтаксических аннотаций. В качестве расширения стандарта ISO 24615-1 этот документ согласован с стандартом ISO 24612.

2.6.2. Структура семантической аннотации

Стандарты на семантическую аннотацию представляют собой самое развитое семейство стандартов, разрабатываемых в ИСО/ТК37/ПК4. В настоящее время это семейство включает 14 стандартов и проектов: Ч.1 Время и события; Ч.2. Диалоговые акты; Ч.4 Семантические роли; Ч.5 Структура дискурса; Ч.6 Принципы семантической аннотации; Ч.7 Пространственная информация; Ч.8 Семантические отношения в дискурсе, ядерная схема аннотаций; Ч.9 Структура справочной аннотации; Ч.10 Визуальная информация; Ч.11 Измеримая количественная информация; Ч.12 Количественная оценка; Ч.14 Пространственная семантика.

Очевидно, что этот перечень далеко не исчерпан.

Приведем примеры аннотаций некоторых стандартов этого семейства.

ISO 24617-1:2012[24] – информация о времени в текстах на естественном языке становится все более важным компонентом понимания этих текстов. Стандарт (SemAF-Time) определяет формализованный язык разметки на основе XML, называемый ISO-TimeML, с систематическим способом извлечения и представления временной информации для облегчения обмена временной информацией как между языковыми системами обработки, так и между различными схемами временного представления.

ISO 24617-2:2020[25] – этот документ содержит набор концепций аннотирования диалога, формальный язык для аннотирования диалога (язык разметки акта диалога, DiAML) и метод сегментации диалога на семантические единицы. Это позволяет вручную или автоматически аннотировать сегменты диалога информацией о коммуникативных действиях, которые участники выполняют в диалоге. В документе поддерживается многомерное аннотирование устных, письменных и мультимодальных диалогов с участием двух или более участников.

ISO 24617-4:2014[26] – представляет собой согласованную схему аннотаций для семантических ролей, т. е. схему, которая указывает на роль, которую участник играет в событии или состоянии, описываемом в основном глаголом. Это включает в себя не только отношения семантические между глаголом и его актантами, но и отношения релевантные для других предикативных элементов, таких как номинализации, существительные, прилагательные и модификаторы предикатов. Предикативная роль наречий выходит за рамки стандарта.

ISO/TS 24617-5:2014[27] – дискурс – это процесс коммуникации. Стандарт рассматривает структуру дис-

[20] ISO 24612:2012 Language resource management – Linguistic annotation framework (LAF). – URL: https://www.iso.org/standard/37326.html

[21] ISO 24611:2012 Language resource management – Morpho-syntactic annotation framework (MAF). – URL: https://www.iso.org/standard/51934.html

[22] ISO 24615-1:2014 Language resource management – Syntactic annotation framework (SynAF) – Part 1: Syntactic. – URL: model https://www.iso.org/standard/62508.html ;ГОСТРИСО 24615-2016 Управление языковыми ресурсами. Система синтаксического аннотирования (SynAF). – URL: http://docs2.kodeks.ru/document/ 1200142745

[23] Language resource management – Syntactic annotation framework (SynAF) – Part 2: XML serialization (Tiger vocabulary). – URL: https://www.iso.org/standard/62491.html

[24] ISO 24617-1:2012 Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events (SemAF-Time, ISO-TimeML). – URL: https://www.iso.org/standard/37331.html

[25] ISO 24617-2:2020 Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts. – URL: https://www.iso.org/standard/76443.html

[26] ISO 24617-4:2014 Language resource management – Semantic annotation framework (SemAF) – Part 4: Semantic roles (SemAF-SR). – URL: https://www.iso.org/standard/56866.html

[27] ISO/TS 24617-5:2014 Language resource management – Semantic annotation framework (SemAF) – Part 5: Discourse

курса с точки зрения его реализации/представления и содержания, а также показывает, как его двойственная структура может быть представлена в виде графика, т. е. определяет аннотации дискурсивных структур только в письменном тексте, но она может быть распространена на дискурсы и в других формах.

ISO 24617-6:2016[28] – определяет структуру семантических аннотаций (SemAF). В нем излагается стратегия SemAF, направленная на долгосрочное объединение отдельных схем аннотаций для определенных классов семантических явлений в единую, согласованную схему семантической аннотации с широким охватом. Стандарт устанавливает понятия абстрактного и конкретного синтаксиса для семантических аннотаций, отражая различие между аннотациями и представлениями в рамках моделей лингвистических аннотаций.

ISO 24617-7:2020[29] – этот документ обеспечивает основу для кодирования широкого спектра пространственной и пространственно-временной информации, относящейся к движению. Документ включает представления местоположений, пространственных объектов, пространственных отношений (включая топологические, ориентационные и метрические значения), события движения, пути, другую пространственную информацию.

ISO 24617-8:2016[30] – устанавливает представление и аннотацию локальных низкоуровневых отношений между ситуациями, упомянутыми в дискурсе, где каждое отношение аннотируется независимо от других отношений в том же дискурсе. Стандарт обеспечивает основу для аннотирования дискурсивных отношений путем указания набора основных дискурсивных отношений. Насколько это возможно документ сопоставляет семантику различных схем.

ISO 24617-9:2019[31] – представляет собой комплексную модель аннотирования и представления референтных явлений в текстах естественного языка и мультимодальных коммуникациях. Такие явления могут включать как простые анафорические или кореферентные, так и более сложные мультимодальные отношения.

В рамках этого семейства в разработке находятся следующие стандарты:

- ISO/AWI 24617-10. Language resource management – Semantic annotation framework (SemAF) – Part10: Visual information (VoxML) Визуальная информация
- ISO/DIS 24617-11. Language resource management – Semantic annotation framework (SemAF) – Part 11: Measurable Quantitative information (MQI) Измеримая количественная информация
- ISO/WD 24617-12. Language resource management – Semantic annotation framework (SemAF) – Part 12: Quantification Количественная оценка
- ISO/AWI 24617-14 Language resource management – Semantic annotation framework (SemAF) – Part 14: Spatial semantics Пространственная информация

2.7. Структура многоязычной информации

ISO 24616:2012 и его российский аналог **ГОСТ Р ИСО 24616-2013**[32] – это универсальная платформа для моделирования и управления многоязычной информацией в различных областях. MILF (Multilingual Information Framework) предоставляет метамодель и набор общих категорий данных для различных областей применения, а также стратегии взаимодействия и/или связывания ЛИР с различными структурами.

Подобно структуре терминологической разметки TMF, многоязычная информационная структура MLIF представляет собой метамодель, которая в сочетании с определенными категориями данных обеспечивает взаимодействие нескольких многоязычных приложений и корпусов.

2.8. Постоянный идентификатор ЛИР

ISO 24619: 2011[33] и его российский аналог **ГОСТ Р ИСО 24619-2013** – определяют требования к структуре постоянного идентификатора (PID) и к его использованию в качестве ссылок на ЛИР в документах, а также в самих ЛИР. Примерами ЛИР являются: цифровые словари, терминологические ресурсы, лексика машинного перевода, аннотированные мультимедийные/мультимодальные или текстовые корпуса, и тому подобные ресурсы.

2.9. Контролируемый естественный язык

ISO/TS 24620-1: 2015[34] – устанавливает принципы контролируемого естественного языка (CNL) и его использования вместе с соответствующей вспомогательной технологией. Однако этот стандарт нацелен на то, чтобы дать общее представление о CNL с его

structure (SemAF-DS). – URL: https://www.iso.org/standard/57083.html

[28] ISO 24617-6:2016 Language resource management – Semantic annotation framework – Part 6: Principles of semantic annotation (SemAF Principles). – URL: https://www.iso.org/ru/standard/60581.html

[29] ISO 24617-7:2020 Language resource management – Semantic annotation framework – Part 7: Spatial information. – URL: https://www.iso.org/standard/76442.html

[30] ISO 24617-8:2016 Language resource management – Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse, core annotation schema (DR-core). – URL: https://www.iso.org/ru/standard/60780.html

[31] ISO 24617-9:2019 Language resource management – Semantic annotation framework – Part 9: Reference annotation framework (RAF). – URL: https://www.iso.org/ru/standard/69658.html

[32] ISO 24616:2012 Language resources management – Multilingual information framework. – URL: https://www.iso.org/standard/37330.html; ГОСТ Р ИСО 24616-2013. Менеджмент языковых ресурсов. Многоязычная информационная система = Language resources management. Multilingual information framework. – URL: https://search.rsl.ru/ru/record/01007837234

[33] ISO 24619:2011 Language resource management – Persistent identification and sustainable access (PISA). – URL: https://www.iso.org/ru/standard/37333.html; ГОСТ Р ИСО 24619-2013 Менеджмент языковых ресурсов. Постоянная идентификация и устойчивый доступ. – URL: http://docs.cntd.ru/document/1200110804

[34] ISO/TS 24620-1:2015 Language resource management – Controlled natural language (CNL) – Part 1: Basic concepts and principles. – URL: https://www.iso.org/ru/standard/37334.html Менеджмент языковых ресурсов. Контролируемый естественный язык (CNL). Часть 1. Общие понятия и принципы. – URL: http://docs.cntd.ru/document/461982398

целями и характеристиками и предоставить схему классификации различных видов CNL.

По данному направлению ПК4 разрабатывает еще ряд документов:

- Контролируемые человеческие коммуникации. Основные принципы и методология контролируемой письменной коммуникации[35].
- Контролируемые человеческие коммуникации. Основные принципы и методология контролируемой устной коммуникации[36].
- Контролируемые человеческие коммуникации. Технические требования к многоязычным текстам[37].

2.10. Инфраструктура компонентов метаданных (CMDI)

ISO 24622:2015[38] – описывает модель, которая позволяет гибко строить схемы взаимодействия метаданных для ЛИР. Схемы метаданных, основанные на этой модели, могут быть использованы для описания ресурсов на различных уровнях детализации Стандарт разделен на 2 части.

ISO 24622-2:2019[39] – стандарт содержит подробные описания и определения того, как выглядят записи CMDI, компоненты и их представления в XML, которые позволяют гибко создавать совместимые схемы метаданных, для описания ЛИР.

2.11. Универсальные языки запросов для корпусов (CQLF)

Этот стандарт состоит из 2-х частей:

ISO 24623-1:2018[40] – описывает абстрактную метамодель, предназначенную для выражения любого языка корпусных запросов. Метамодель состоит из нескольких компонентов, называемых классами, уровнями и модулями CQLF. В данном документе рассматриваются три уровня CQLF (линейный, сложный и параллельный), а также их разделение на модули, продиктованные функциональными и другими критериями.

ISO/DIS 24623-2[41] – представляет онтологию CQLF, которая определяет выразительную силу языков корпусных запросов, поддерживаемых операторами соответствия в форме параметризованных выражений запросов. На основе этого проекта стандарта онтология будет создаваться, обновляться и расширяться в рамках процесса, совместно управляемого разработчиками языков корпусных запросов и конечными пользователями. Этот процесс модерируется организацией CQLF Ontology GitHub. Онтология использует идентификатор пространства имен, доступный по адресу: https://www.clarin.eu/standards/cqlf.

2.12. Транскрипция устной речи

ISO 24624: 2016[42] – определяет правила представления транскрипций аудио- и видеозаписанных речевых взаимодействий в XML-документах на основе руководящих принципов TEI. Вторичная цель документа: связать транскрибированные данные со стандартами для аннотированных корпусов.

ЗАКЛЮЧЕНИЕ

Из представленного в настоящей статье обзора следует, что комитеты ИСО/ТК37/ПК3 и особенно ИСО/ТК37/ПК4 проводят большую и содержательную работу по анализу и обобщению различных лингвистических информационных ресурсов, обеспечивая тем самым возможность их интеграции и повторного использования. Результат этой работы весьма впечатляет. Однако с реальным внедрением разработанных стандартов дело обстоит менее убедительно. Из многочисленных международных сообществ, связанных с ЛИР, пожалуй, только CLARIN активно использует разработанные стандарты ИСО.

Особенно наглядно этот недостаток деятельности по стандартизации ЛИР проявляется в России. Действующий в настоящее время Технический комитет 55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле» перевел несколько стандартов по управлению ЛИР из числа разработанных в ИСО (сведения о наличии российских аналогов приводятся в настоящем обзоре) и утвердил их в качестве национальных стандартов. Выбор стандартов для перевода производит, впрочем, впечатление случайного, а качество переводов при этом чрезвычайно низкое: кажется, что результаты автоматического перевода вообще не редактировались.

Главный же недостаток деятельности российского ТК55, заключается в том, что разработанные им стандарты вообще не применяются при разработке российских лингвистических информационных ресурсов. Это неудивительно, ведь в составе этого ТК практически нет разработчиков ЛИР. Исключение только одно – возглавляет ТК55 Стандартинформ, который поддерживает известный российский банк терминологических данных РОСТЕРМ. Однако этот банк данных не замечен в активном сотрудничестве с другими разработчиками отечественных ЛИР, а также в разработке открытого доступа и интеграции терминологических данных.

В данной ситуации руководству Росстандарта можно порекомендовать больше обращать внимание на

[35] ISO/WD 24620-2 Language resource management – Controlled human communication (CHC) – Part 2: Basic principles and methodology for controlled written communication (CWC). – URL: https://www.iso.org/standard/74581.html

[36] ISO/FDIS 24620-3 Language resource management – Controlled human communication (CHC) – Part 3: Basic principles and methodology for controlled oral communication (COral-Com). – URL: https://www.iso.org/standard/76446.html

[37] ISO/WD 24620-4 Language resource management – Controlled human communication (CHC) – Part 4: Multilingual technical requirements. – URL: https://www.iso.org/ru/standard/79087.html

[38] ISO 24622-1:2015 Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. – URL: https://www.iso.org/ru/standard/37336.html

[39] ISO 24622-2:2019 Language resource management – Component metadata infrasctructure (CMDI) – Part 2: Component metadata specification language. – URL: https://www.iso.org/standard/64579.html

[40] ISO 24623-1:2018 Language resource management – Corpus query lingua franca (CQLF) – Part 1: Metamodel. – URL: https://www.iso.org/standard/64579.html

[41] ISO/DIS 24623-2 Language resource management – Corpus Query Lingua Franca (CQLF) – Part 2: Ontology

[42] ISO 24624:2016 Language resource management – Transcription of spoken language – URL: https://www.iso.org/standard/37338.html

внедрение национальных стандартов, разработанных ТК55 «Терминология, элементы данных и документация в бизнес-процессах и электронной торговле».

СПИСОК ЛИТЕРАТУРЫ

1. Romary L. An abstract model for the representation of multilingual terminological data: TMF – Terminological Markup Framework // Laboratoire LORIA Campus Scientifique BP 239, F-54506 Vandoeuvrelès-Nancy. – URL: http://www.termsciences.fr/IMG/pdf/ Romary2.pdf
2. The Standards' Landscape Towards an Interoperability Framework The FLaReNet proposal Building on the CLARIN Standardisation Action Plan July 2011. – URL: http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf
3. Standards for language resources in ISO – Looking back at 13 fruitful years Laurent Romary. – URL: https://arxiv.org/ftp/arxiv/papers/1510/1510.07851.pdf
4. Standards for Language Resources /Nancy Ide, Laurent Romary. – URL: https://www.researchgate.net/ publication/301865825_Standards_for_Language_Resources
5. The Standards' Landscape Towards an Interoperability Framework. The FLaReNet proposal Building on the CLARIN Standardisation Action Plan July 2011. – URL: http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf
6. The Text Encoding Initiative (TEI) Инициатива по кодированию текстов. – URL: https://tei-c.org/
7. The TEI Guidelines for Electronic Text Encoding and Interchange. – URL: https://tei-c.org/guidelines/
8. Захаров В.П. Международные стандарты в области корпусной лингвистики // Структурная и прикладная лингвистика. – 2012. – № 9. – С. 201-221
9. Савчук С.О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. – С. 62-88.
10. Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Sori. Lexical Markup Framework (LMF). – URL: https://www.academia.edu/13918909/Lexical_markup_framework_LMF_

Материал поступил в редакцию 10.02.21.

Сведения об авторе

АНТОПОЛЬСКИЙ Александр Борисович – доктор технических наук, профессор, главный научный сотрудник ИНИОН РАН, Москва
e-mail: ale5695@yandex.ru