УДК 81’42 : 81’33

В.С. Селезнева
Белорусский государственный университет, Минск, Республика Беларусь

V.S. Selezneva
Belarusian State University, Minsk, Republic of Belarus

ФОРМАЛИЗАЦИЯ НАУЧНЫХ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

FORMALISATION OF SCIENTIFIC RUSSIAN-LANGUAGE TEXTS

Рассматриваются особенности научных русскоязычных текстов и возможные способы их формализации, что поспособствует адаптации таких текстов для дальнейшего использования в условиях электронной информационной среды. Выделяется проблематика, с которой сталкиваются при обработке сложных синтаксических конструкций таких текстов посредством компьютерной обработки. Отмечается, что научные русскоязычные тексты в электронной среде имеют перспективу для широкого применения.

Ключевые слова: формализация; научный текст; токенизация; лемматизация; метаанализ.

The features of scientific Russian-language texts and possible ways of their formalisation are considered, that will promote the adaptation of such texts for further usage in the conditions of the electronic information environment. The problems which people face while processing complex syntactic constructions of such texts by means of computer processing are highlighted. It is noted that scientific Russian-language texts in the electronic environment have a prospect for a wide application.

Key words: formalisation; scientific text; tokenisation; lemmatisation; meta-analysis.

Под формализацией научных русскоязычных текстов понимается процесс преобразования естественного языка [Мельчук 2012] научных работ в строгие, структурированные формы. Этот процесс имеет свои особенности, связанные как с общими принципами научной коммуникации, так и с лингвистическими и культурными аспектами русскоязычной научной традиции. Рассмотрим основные аспекты данной проблемы.

«Как известно, научный текст отличается от других набором языковых средств, структурной организацией, нормами употребления лексических, грамматических и синтаксических построений» [Романенко 2011: 353]. Для русскоязычных научных текстов характерно использование сложных синтаксических конструкций (например: причастных/деепричастных оборотов). Также научные тексты на русском языке часто содержат специализированную терминологию, которая может быть многозначной или иметь аналоги в других языках. Следует отметить, что в русской научной традиции акцент может делаться на теоретические аспекты, что отражается в структуре текста (например, больше внимания уделяется обоснованию, чем практическим результатам).

В ходе формализации научных текстов на русском языке могут возникнуть следующие проблемы: многозначность терминов, так, один и тот же термин может иметь разные значения в разных контекстах; отсутствие четких стандартов в связи с тем, что в некоторых областях науки отсутствуют унифицированные подходы к описанию данных; сложности обработки сложных конструкций, так как длинные предложения и вводные конструкции затрудняют автоматический анализ текста; проблемы интеграции с международными системами по причине неполного соответствия русскоязычных терминов при переводе на английский язык и их сопоставлении с международными базами данных.

В качестве методов формализации могут быть применены лингвистический анализ, заключающийся в токенизации (разделение текста на слова и предложения) и лемматизации (приведение слов к начальной форме) [Пименов, Воронов http], а также извлечение терминов и ключевых фраз может производиться с помощью семантического анализа, включающего в себя выявление связей между понятиями и классификацию научных текстов по тематике. В ходе формализации проводится разделение текста на разделы (введение, методы, результаты, обсуждение) и извлечение данных из текста (например, числовых значений, формул, таблиц), при этом происходит создание моделей той или иной предметной области и сопоставление терминов с общемировыми.

Процесс формализации структуры научных русскоязычных текстов способствует эффективной обработке таких текстов в электронной информационной среде. В данном случае подразумевается автоматическая обработка больших объемов текстов естественного языка посредством компьютерных технологий.

В качестве инструментов и технологий по обработке и формализации научных русскоязычных текстов могут быть использованы: NLP-библиотеки: например, spaCy, NLTK, Stanza, которые поддерживают обработку русского языка; машинное обучение: использование моделей для классификации текстов, извлечения информации и анализа тональности; тезаурусы: например, RuThes (русский тезаурус) или интеграция с WordNet; системы управления знаниями: например, Protégé для создания онтологий.

Сфера применения форматизированных данных по русскоязычным научным текстам может быть самой широкой: анализ научных публикаций, при котором происходит автоматическое извлечение данных из статей для метаанализа; улучшение поиска по научным базам данных посредством различных поисковых систем (например: eLibrary, CyberLeninka); формализация текстов для использования в экспертных системах; адаптация научных русскоязычных текстов для международной аудитории.

В перспективе все это неизбежно приведет к развитию NLP-технологий для русского языка, использование искусственного интеллекта для автоматической формализации и анализа текстов, а также интеграции русскоязычных научных текстов в соответствии с международными стандартами (например, DOI, ORCID), что открывает для них более широкую аудиторию.

Формализация научных русскоязычных текстов – это важный шаг к повышению доступности и интеграции российских исследований в мировую научную среду. Однако для успешной реализации требуется учет лингвистических и культурных особенностей, а также развитие соответствующих технологий.

Список литературы

Мельчук И. А. Язык: от смысла к тексту. М. : Языки славянской культуры (ЯСК), 2012.

Пименов В. И., Воронов М. В. Формализация регулятивных текстов. URL: https://ia.spcras.ru/index.php/sp/article/view/14344 (дата обращения: 20.01.2025).

Романенко О. Н. Применение метода формализации научного текста на занятиях по русскому языку как иностранному // Вестн. ТГГПУ. 2011. № 4 (26). С. 353–356.