Лясович Светлана Михайловна
Полоцкий государственный университет
имени Евфросинии Полоцкой
e-mail:s.liasovich@psu.by

ОПЫТ СОЗДАНИЯ ЛИНГВИСТИЧЕСКОГО КОРПУСА SEAH И ОСНОВЫ РАБОТЫ С НИМ

Аннотация. В статье описан опыт создания профессионально ориентированного русскоязычного корпуса текстов для студентов направления «Архитектура и строительство» с использованием корпусного менеджера Sketch Engine в рамках проекта SEAH. Обозначены цели и задачи проекта, подход к отбору материала для корпуса, обоснована сбалансированность корпуса, включение в корпус текстов разных жанров и стилей. Описаны элементы поисковой системы в связи с разметкой лингвистического корпуса.

Ключевые слова: корпусная лингвистика, лингвистический корпус, SEAH, Sketch Engine, Национальный корпус русского языка, профессиональный языковой модуль.

Sviatlana Liasovich
Euphrosyne Polotskaya State University of Polotsk
e-mail: s.liasovich@psu.by

EXPERIENCE OF CREATING THE LINGUISTIC CORPORA SEAH AND THE BASIS OF WORKING WITH IT

Abstract. The article describes the experience of creating a professionally oriented Russian-language text corpus for students of Architecture and Construction using the Sketch Engine corpus manager as part of the SEAH project. The goals and objectives of the project, the approach to the selection of material for the corpus, the balanced nature of the corpus, and the inclusion of texts of different genres and styles in the corpus are outlined.It also deals with the elements of the search engine in relation to the annotation of the text corpus.

Keywords: corpus linguistics, linguistic corpus, SEAH, Sketch Engine, Russian national corpus, professional language module.

В настоящее время корпусная лингвистика видится актуальным направлением филологических исследований как при изучении явлений языка, так и при обучении языку. Созданные лингвистические корпуса изменили и взгляд на язык, и инструментарий исследователя. Как отмечает В. П. Захаров, «под названием

лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [1, с. 3]. Кроме лингвистических задач, сегодня созданные лингвистические корпуса используют и для решения ряда методических задач.

Целью данной статьи является представление части результатов проекта «Совместное использование европейского архитектурного наследия: инновационные инструменты преподавания языков для академической и профессиональной мобильности по направлению «Архитектура и строительство», над которым работает кафедра славянской филологии Полоцкого государственного университета имени Евфросинии Полоцкой. Проект SEAH реализуется в рамках инструмента европейского стратегического партнерства программы Эразмус+ и направлен на активизацию программы академической мобильности студентов, которые изучают архитектуру и строительство, в первую очередь, стран, которые являются участниками проекта. Это Италия, Чехия, Франция, Испания и Беларусь (сайт проекта https://www.seahproject.eu/).

Основная цель проекта – научно-исследовательская деятельность для инноваций в области образования.

Проект предполагает создание профессиональных языковых модулей для французского, немецкого, итальянского, русского и испанского языков как иностранных в области архитектуры и строительства. Можно выделить 2 основных этапа работы: 1) создание языковых корпусов, включающих тексты, отражающие профессиональное и академическое общение в области архитектуры и строительства; 2) каждый университет-партнер разрабатывает модули онлайн-обучения языку в области архитектуры и строительства.

В данной статье рассмотрим работу первого этапа, во время которого совместно с коллегами из Университета Д'Аннунцио Кьети-Пескара (Италия) был создан корпус текстов на русском языке. Университет Бордо Монтень (Франция) разрабатывал корпус текстов на французском языке, Мазарик Университет (Чешская Республика) – на немецком языке, Политехнический Университет (г. Мадрид, Испания) – на испанском. Размещением всех материалов на специальной платформе занималась компания Web Solutions (г. Малага, Испания).

Созданные лингвистические корпуса профессионально ориентированные на сферу архитектуры и строительства. Научная разработанность проблемы профессионально ориентированного обучения русскому языку как иностранному, по нашему мнению, представляется недостаточной, особенно области архитектуры и строительства. Профессионально ориентированное обучение языку «направляет педагогический процесс на конечный результат обучения студента в вузе –

будущую профессию, которая в итоге станет сферой приложения всех получаемых знаний, умений, навыков, проверкой их действенности» [2, с. 3]. В этом специфика такого обучения. Но, как отмечает И. А. Пугачев, профессионально ориентированное обучение русскому языку «не может быть сведено только к изучению языка специальности и тем более только к изучению научного стиля речи, хотя мы и признаем их ведущими аспектами профессионально ориентированного обучения. Сам термин «профессионально ориентированное обучение» используется нами для обозначения всего процесса организации преподавания русского языка в неязыковом вузе, включающего обучение иностранных учащихся русскому языку и как средству получения высшего образования и специальности, и как средству общения в диалоге культур, и как средству гуманистического развития учащихся средствами русского языка» [5, с. 38]. Поэтому в поле зрения составителей корпуса вошли и тексты публицистического, научно-популярного, официально-делового стилей.

По ссылке https://corpora.unich.it/seah/#dashboard?corpname=ru_seah можно познакомиться с составленным корпусом текстов на русском языке. Что касается платформы, на которой размещались лингвистические корпуса, то это ресурс Sketch Engine. «В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют *корпусным менеджером* (или корпус-менеджером) (*англ*. corpus manager). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме» [1, с. 3] Sketch Engine и является таким корпусным менеджером.

Исследователь может сам создавать корпус на этом ресурсе, для чего необходима регистрация. Для России и Беларуси в данный момент этот ресурс недоступен для моделирования собственных корпусов, но размещение собранных текстов, техническую обработку осуществляла компания Web Solutions. Поэтому в качестве пользователей работа с ресурсом возможна.

Зарегистрированный пользователь может загрузить заранее подготовленный текстовый корпус объемом не более 1 миллион слов на одном из языков, которые поддерживает Sketch Engine. Подробную информацию о работе с данным корпусным менеджером, шаблонов поиска можно найти в руководстве пользователя (*User guide*).

Под электронным корпусом в статье понимается «собрание текстов на данном языке, представленное в электронном виде и снабжённое научным аппаратом (разметкой)» [4, с. 6]. Обе названные характеристики – электронная форма и наличие разметки – основное, что отличает электронные текстовые корпусы, с одной стороны, от традиционных текстовых корпусов, с другой стороны, от любого

собрания текстов в электронной форме. Разметка представляет собой ту информацию, которая вносится в тексты при их обработке, и зависит от цели, с которой составляется корпус.

К первичной разметке, присутствующей практически в каждом корпусе, относят токенизацию (разбиение на орфографические слова), лемматизацию (приведение словоформ к словарной форме) и парсинг (синтаксический анализ). Подробнее этот вопрос освещен в работе Е. Б. Кротовой [3].

Кроме того, необходимо обращать внимание на состав корпуса. Для ряда исследований важно, чтобы корпус был сбалансированным. С помощью сбалансированного корпуса можно получить представление о том, как изучаемый языковой феномен проявляет себя в целом в языке. Сбалансированным считается корпус, который «содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода» [6]. Синонимичный термин, употребляемый В. П. Захаровым, – репрезентативность. «Под *репрезентативностью* понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т. п.» [1, с. 5].

Что касается созданного корпуса текстов по архитектуре и строительству, то стремление сделать его сбалансированным, обусловило включение в корпус текстов разной направленности и жанров. По области применения текстов выделяются академические, профессиональные, педагогические и популярные. В область академических входят такие жанры как научные статьи, монографии, материалы конференций, авторефераты диссертаций, диссертации и др.; в область профессиональных – описание проектов, отчеты по проектам, в сферу педагогических – пособия, лекции, учебные материалы, методички, в область популярных – публичные лекции, подкасты, буклеты торговых выставок и др. Эта информация представлена в виде метаразметки. «Под метаразметкой понимается приписывание тексту атрибутов, характеризующих обстоятельства его создания, автора, тематику, жанровые особенности и др.» [6]. Эта информация облегчает целевой поиск по корпусу, а также делает собрание текстов удобным для использования не только методистами-филологами, но и для тех, кто изучает сферу архитектуры и строительства.

Изучение языка связано и с освоением навыка понимания на слух, поэтому включение в корпус как письменных, так и устных текстов также отражает стремление сделать его сбалансированным. Материалом устных текстов по большей части стали записи лекций преподавателей Полоцкого государственного университета имени Евфросинии Полоцкой, записи защиты дипломных и курсовых

проектов студентов специальности «Архитектура», записи выступлений с докладами на научных конференциях.

Большинство текстов, размещенных в корпусе, соответствует сертификационному уровню русского языка B1, B2. Кроме того, нужно отметить, что материал может быть интересен специалистам в сфере строительства и архитектуры из других стран, т. к. представляет собой местные региональные особенности развития этой области. Таким образом, созданный ресурс можно рассматривать как платформу для работы не только лингвистов и методистов, но и как объект междисциплинарных связей.

СПИСОК ЛИТЕРАТУРЫ

1. Захаров, В. П. Корпусная лингвистика: Учебно-метод. пособие / В. П. Захаров. – СПб. : Санкт-Петербургский государственный университет, 2005. – 48 с.
2. Коренева, А. В. Профессионально ориентированное обучение речевой деятельности студентов-нефилологов на основе междисциплинарной интеграции: автореф. дис. … канд. пед. наук : 13.00.02 / А. В. Коренева. – Орел, 2009. – 41 с.
3. Кротова, Е. Б. Омонимия и проблемы разметки электронных корпусов / Е. Б. Кротова // Грамматические категории германских языков в антропоцентрической перспективе. Коллективная монография / Отв. ред. Д. Б. Никуличева. Ред. кол.: Н. С. Бабенко, В. А. Нуриев, В. И. Карпов, Е. Б. Кротова, Т. В. Топорова, Е. Б. Яковенко. – М. : Канцлер, 2017. – С. 248–256.
4. Плунгян, В. А. Зачем нужен Национальный корпус русского языка? / В. А. Плунгян // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М. : Индрик, 2005. – С. 6–21.
5. Пугачев, И. А. Профессионально ориентированное обучение русскому языку как иностранному: теория, практика, технологии : монография / И. А. Пугачев. – Москва : РУДН, 2016. – 483 с.
6. Что такое корпус? / НКРЯ. – Режим доступа: https://ruscorpora.ru/old/corpora-intro.html. – Дата доступа: 24.06.2022.