Учебные корпуса содержат коллекцию текстов письменной или затранскрибированной устной речи, порожденной людьми, изучающими язык. Основная задача в рамках создания учебных корпусов состоит в нахождении и разметке ошибок, допущенных изучающими язык. Это означает, что, в отличие от стандартных языковых корпусов, в учебных корпусах применяется система тэгов, определяющих типы ошибок, а не система тэгов с грамматической характеристикой слов. К ним могут относиться ошибки выбора и использования слов, словообразовательные, грамматические и морфосинтаксические ошибки. Полученные данные затем анализируются на предмет межъязыковых сходств и различий в усвоении различных аспектов изучаемого языка.

Учебные корпуса также справляются с недостатками и ограничениями традиционного анализа ошибок (Error Analysis). Впервые идея анализа ошибок при усвоении второго языка возникла в 1970 году, до этого момента ошибки рассматривались как нечто нежелательное, но неизбежно возникающее при изучении иностранных языков (Ellis, 1994). Позже возникла идея, что анализ ошибок позволяет описать владение языком изучающими, помогая тем самым учителям, исследователям и самим обучающимся (Corder, 1981). Учителя получают информацию о степени успешности усвоения системы языка и о том, каким аспектам следует уделить больше внимания в процессе обучения. Исследователи обнаруживают стратегии и методы, используемые студентами в процессе обучения. Обучающиеся же получают возможность увидеть допускаемые ими и другими студентами типы ошибок. В работе (Ellis, 1994) выделяется четыре этапа анализа ошибок: обнаружение ошибки, ее описание, объяснение и оценка. В результате перехода от немашинного анализа ошибок к корпусным исследованиям удалось создать удобную систему, пригодную для обработки больших баз данных, выделить большинство возможных ошибок, а также проводить статистические подсчеты.

Исходя из вышесказанного, цель моей работы заключается в изучении ответов, русских студентов в устной части экзамена IELTS, предпринимаемом для создания системы аннотации, которая позволяла бы размечать типичные и особые ошибки, совершаемые изучающими английский язык, а также получать сведения о частоте различных видов ошибок. Конечной целью является создание аннотированного учебного корпуса, содержащего коллекцию транскрибированных текстов, порожденных носителями русского языка, изучающими английский язык. Разработка корпуса послужит основой для дальнейшего количественного и качественного исследования устных ответов. В то время как существует множество учебных корпусов, отличительным признаком данного корпуса видится сосредоточение внимания исключительно на ошибках русских студентов, изучающих английский как иностранный, что позволит в деталях рассмотреть особенности усвоения английского языка при учете влияния родного языка изучающих - русского. Кроме того, необходимо скорректировать систему тэгов в связи с особенностями устной речи. Как будет продемонстрировано в дальнейшем, имеется мало сведений об изучении устной компетенции русскоговорящих студентов в английском языке.

Кембриджский учебный корпус содержит коллекцию письменных ответов на кембриджские экзаменационные задания на знание английского языка (KET, PET, FCE, CAE, CPE, IELTS и другие). Корпус содержит тексты, написанные более чем 180000 студентами из 200 стран и говорящих на 86 языках в качестве родного языка. Информация о каждом студенте включает в себя сведения о первом языке, возрасте, поле, истории изучении языка. Эти сведения дают возможность разделить корпус на подкорпуса в зависимости, например, от возраста или уровня владения языком. В настоящее время корпус содержит 6 миллионов размеченных слов, при этом работа выполнялась лишь двумя людьми, что позволило избежать непоследовательности в выборе тэгов. Как и в японском корпусе, создатели обращают внимание на необходимость лишь аннотации и исправления ошибок, но не переписывания предложений полностью с целью создать совершенно правильные предложения.

Корпусные данные используются издательством Кембриджского университета в

лексикографии для определения слов и конструкций, вызывающих наибольшие затруднения у изучающих английский язык.

Японский учебный корпус (The NICT JLE Corpus) был создан в 2004 году национальным институтом информационных и коммуникационных технологий. Источником данных являются транскрибированные аудиозаписи, записанные во время экзамена The Standard Speaking Test. Данное тестирование было разработано в Японии и направлено на оценку уровня устного владения английским языком. Авторы заявляют, что экзамен может служить альтернативой стандартным письменным тестам.

Корпус содержит 1281 документ и 1.2 миллиона слов. В японском учебном корпусе существует два варианта разметки: основные тэги (для всех документов) и тэги ошибок (для части документов). Основные тэги включают в себя структуру интервью, тэги профиля интервьюера, тэги различных речевых явлений: филлеры, повторения, самоисправления, наложения, и другие. Система тэгов, описывающих ошибки, охватывает грамматические и лексические ошибки и содержит 47 тэгов.

К особенностям японского учебного корпуса можно отнести тот факт, что сведения о говорящем содержат уровень владения языком по девятибалльной шкале. Это позволяет проводить анализ и сравнения внутри выбранного языкового уровня. Кроме того, в корпусе существует подкорпус текстов, порожденных носителями языка, выполнявших аналогичные задания. Подкорпус дает возможность сравнивать речь носителей английского и носителей японского, которые изучают английский язык.

Создатели корпуса отмечают сложность создания системы тэгов, которые охватывали бы все возможные виды ошибок, поскольку ошибки изучающих язык сильно варьируются, и включают в себя не только грамматические и лексические ошибки, но и неправильное произношение, дискурсивные ошибки и другие типы ошибок, связанные с коммуникативной составляющей языка. В созданной авторами системе тэги написаны в формате XML и состоят из трех частей: часть речи, морфологическое, лексическое или грамматическое правило, правильная форма. Сложности также возникают в процессе разметки ошибок, так как ошибки могут накладываться друг на друга, или одной ошибке может быть приписано несколько тегов. Для решения трудностей, связанных с разметкой ошибок, был выработан ряд правил, основной принцип которых – исправление ошибок с минимально возможным изменением структуры предложения (Izumi, и др.).

В настоящее время ведется разработка ланкастерского устного корпуса. Данный корпус является совместным проектом CASS Business School и лондонского Тринити-колледжа (международная площадка для проведения экзаменов, в том числе международных экзаменов по английскому языку (ESOL)). Цель проекта – создание большого корпуса устной речи на основе экзаменационных ответов студентов, изучающих английский язык. Корпус предполагается использовать для различных исследовательских направлений, таких как усвоение второго языка, тестирование уровня владения иностранным языком, преподавание второго языка, составление учебных материалов и других. Корпус составлен на основе устных ответов экзамена GESE, уровень владения варьируется от уровня B1 до уровня C2 по европейской системе уровней владения языком. Особое внимание в экзамене уделяется коммуникационным навыкам, где экзаменуемый активно взаимодействует с экзаменатором. Экзаменующиеся являлись носителями итальянского, испанского, китайского, индийского и русского языков из 9 стран. Возраст отвечающих варьируется от 9 до 72 лет. К настоящему моменту корпус содержит 3000000 слов, данные разделены на 3 уровня владения языком. Свободный доступ к материалам корпуса планируется открыть в 2017 году.