МИНИСТЕРСТВО ОБРАЗОВАНИЯ НАУКИ

РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Тверской государственный университет»

Факультет иностранных языков и международной коммуникации

Направление «Лингвистика»

Кафедра английского языка

РЕФЕРАТ

на тему: «Аннотированные корпуса»

Выполнила:

студентка 1 курса группы 13

дневного отделения

Кулакова Полина Дмитриевна

 Проверила:

к.ф.н., доцент,

доцент кафедры английского языка

Е.М. Масленникова

ТВЕРЬ

2021

**Содержание:**

|  |  |
| --- | --- |
| Введение ………………………………………………………………………. | 3 |
| Русские аннотированные корпуса …………………………………………... | 3 |
| Хельсинкский аннотированный корпус (ХАНКО) ………………………… | 4 |
| Тюбингенский корпус (ТК)…………………………………………………... | 5 |
| Корпус газетных текстов (КГТ)……………………………………………… | 5 |
| Заключение …………………………………………………………………… | 6 |
| Список литературы …………………………………………………………... | 7 |

**Введение**

На протяжении многих лет русский язык оставался одним из немногих крупнейших языков, не имеющих собственного корпуса текстов, доступного для исследователей всего мира. Между тем в последнее время, когда появление корпусов привело к существенным изменениям исследовательских практик, если не к серьезным сдвигам в осознании феномена языка [Fillmore 1992], необходимость создания корпуса русских текстов была ясно осознана как российскими, так и зарубежными лингвистами. Неудивительно, что почти одновременно в разных странах возникли научные проекты по разработке корпусов русского языка для размещения в Интернете.

В 1999 г. в университете г. Тюбинген (Германия) началась работа над корпусом, в основу которого лег Уппсальский корпус русских текстов. В Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ с 2000 г. создается Корпус текстов русских газет конца ХХ века. Через год на Отделении славянских и балтийских языков и литератур Хельсинкского университета началась разработка Хельсинкского аннотированного корпуса ХАНКО. Почти одновременно в рамках программы РАН «Филология и информатика» стартовал проект по созданию Национального корпуса русского языка.

Конечно, существующие русские корпуса не исчерпываются вышеперечисленными ресурсами. Самым известным корпусом русского языка долгие годы оставался уже упомянутый Уппсальский корпус, созданный шведскими русистами. В первоначальном виде его объем составлял 1 млн. словоупотреблений, в нем отсутствовала лемматизация и морфологическая разметка. Тем самым ни по размеру, ни по составу информации, которой снабжены тексты, Уппсальский корпус, к сожалению, не отвечает современным стандартам составления корпусов. Во многом именно осознание его недостаточности для адекватного представления языка, ограниченной применимости для лингвистических исследований, а также устарелости материалов привело к активизации работ по созданию альтернативных корпусов. Из других проектов, предшествовавших эре интернет-корпусов, следует упомянуть Машинный фонд русского языка, создававшийся с конца 1980-х гг. под руководством В. М. Андрющенко и А. П. Ершова [Андрющенко 1989]. Хотя проект не был завершен и не привел к разработке представительного корпуса, в его рамках были собраны коллекции текстов разного типа. К проектам недавнего времени относятся корпус текстов по современной публицистике (1990-е гг.), создававшийся в Отделе экспериментально лексикографии Института русского языка А. Н. Барановым, М. Н. Михайловым и Г. О. Сидоровым [1998], параллельный русско-финский корпус в университете г. Тампере в. Финляндии, разрабатываемый М. Н. Михайловым и Х. Томмолой [Михайлов 2003], а также корпус СПБГУ, создаваемый под руководством В. Б. Касевича [Венцов, Касевич 1998]. Указанные корпуса не являются доступными для широкого круга пользователей, информацию об их исследовательском потенциале, их характеристики можно извлечь пока только из созданных разработчиками описаний.

В реферате рассматриваются существующие общедоступные корпуса современного русского языка, сведения о которых, как мне кажется, представляют наибольший практический интерес для исследователей языка и всех интересующихся теми или иными аспектами его функционирования (в реферате не учитываются 2 диахронических текста, которые, в силу своих особенностей, обладают целым рядом отличий).

Кроме НКРЯ, это следующие корпуса:

|  |  |
| --- | --- |
| Хельсинский аннотированный корпус (ХАНКО) | http www. ling.helsinki.fi/projectsIhancol |
| Тюбингенский корпус русского языка (ТК) | hipyiwww.stbaTunituebingen.de/b1/rus/korpora.html |
| Корпус газетных текстов русского языка конца XX века (КГТ) | https www philol msu.ral~Texlcorpus |

Рассмотрим каждый корпус подробнее.

**Хельсинкский аннотированный корпус (ХАНКО)**

Корпус ХАНКО задуман как составная часть проекта «Функциональный синтаксис русского языка», работа над которым ведется на Отделении славянских и балтийских языков и литератур Хельсинкского университета. Постановка задачи предполагает необходимость внесения в корпус подробных лингвистических данных о входящих в его состав единицах и, соответственно, в значительной степени требует ручной обработки текстов. Это определило один из основных принципов построения корпуса — его направленность на максимальный охват грамматической информации, а не на объем материала.

Проектом по созданию корпуса ХАНКО руководит профессор Отделения славистики и балтистики А. Мустайоки, главным разработчиком корпуса является М. Копотев. В работе принимают участие сотрудники Хельсинкского и Тартуского университетов, а также русскоязычные студенты .

В настоящее время осуществление проекта продолжается. В интернете доступны результаты первого этапа работы — морфологически аннотированный корпус. В ближайшее время появится и синтаксическая разметка.

**Тюбингенский корпус (ТК)**

Тюбингенский корпус русского языка создавался в рамках междисциплинарной исследовательской программы, посвященной теоретическим и эмпирическим основам грамматического исследования («Linguistische Datenstrukturen: Theoretische und empirische Grundlagen der Grammatikforschung»). В проекте участвуют различные исследовательские группы, каждая из которых занимается решением конкретных научных задач, связанных с общей тематикой программы. Одна из таких групп с 1999 по 2004 г. работала над описанием форм обращения и вежливости в славянских языках, прежде всего в русском и чешском, на основе корпусов этих языков. К началу проекта корпусной анализ языковых явлений чешского языка представлялся вполне выполнимой задачей благодаря существованию Чешского национального корпуса — одного из первых по времени создания представительных корпусов славянских языков. Для русского языка аналогичная задача оказалась более трудоемкой: из-за отсутствия корпуса, который бы отражал современное языковое употребление, изучение того или иного явления требовало предварительного создания базы для исследования. Тем самым построение по возможности большого и представительного корпуса русского языка стало одной из целей проекта.

Общее руководство проектом осуществлял профессор Отделения славистики университета г. Тюбинген Т. Бергер, в проекте приняли участие преподаватели и аспиранты отделения. ТК стал первым корпусом русского языка, появившимся в открытом доступе в сети Интернет.

В настоящее время проект по исследованию форм обращения и вежливости и, соответственно, работа над корпусом завершены.

**Корпус газетных текстов (КГТ)**

«Компьютерный корпус газетных текстов русского языка конца ХХ века» был подготовлен в течение 2000-2002 гг. в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ.

"Создание корпусов и других коллекций текстов является одним из основных направлений деятельности лаборатории. Здесь в разное время разрабатывались, например, корпус русских литературно-художественных текстов ХИХ-ХХ вв., корпуса русских рассказов, психологических самоописаний, сочинений иностранцев на русском языке, параллельных текстов на русском и английском языках. В отличие от газетного корпуса, все прочие текстовые коллекции не являются общедоступными.

Проект по созданию корпуса газетных текстов осуществляется под руководством заведующего Лабораторией общей и компьютерной лексикологии и лексикографии А. А. Поликарпова, в разработке корпуса принимают участие сотрудники Лаборатории [Винотрадова и др. 2001].

В настоящее время в Интернете доступен небольшой тестовый фрагмент корпуса, более полная версия готовится к представлению.

Хочется вновь упомянуть, что под лингвистической аннотацией или разметкой корпуса (по-английски linguistic markup) подразумевается наличие в корпусе неких данных, не являющихся частью текста, но несущих какую-то информацию о нём (так называемые метаданные). Простейший пример таких данных — отметки частей речи. Выглядеть это может так:

I will use Google before asking dumb questions.

Размечаем:

I (pronoun)will (verb) use (verb) Google (noun) before (preposition) asking (verb) dumb (adjective) questions (noun).

Это нужно для облегчения автоматического анализа корпуса. Один раз отметив в тексте все части речи, затем можно производить любые исследования, связанные с ними без необходимости заново выявлять, например, все прилагательные в корпусе. Понятно, что если такой разметки нет, то, к примеру, поиск по слову «will» выдавал бы все случаи его появления в корпусе, вне зависимости от того, существительное это или вспомогательный глагол.

Это далеко не единственный тип разметки, который бывает нужен корпусному лингвисту, именно поэтому аннотированные корпуса так важны, прошу простить мою тавтологию, корпусным лингвистам.

**Список литературы:**

Тематический портал «Разумное. Доброе. Вечное» <https://myfilology.ru/177/korpusy-annotirovannye-i-neannotirovannye-lingvisticheskaya-annotacziya-razmetka-i-metadannye/#:~:text=%D0%BC%D0%B0%D0%BD%D1%83%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%82%D0%BE%D0%B2%20%E2%80%94%20%D1%8D%D1%82%D0%BE%20%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%82%D0%BA%D0%B0.-,%D0%9F%D0%BE%D0%B4%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9%20%D0%B0%D0%BD%D0%BD%D0%BE%D1%82%D0%B0%D1%86%D0%B8%D0%B5%D0%B9%20%D0%B8%D0%BB%D0%B8%20%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%82%D0%BA%D0%BE%D0%B9%20%D0%BA%D0%BE%D1%80%D0%BF%D1%83%D1%81%D0%B0%20> (Дата обращения: 01.06.2021)

<https://ruscorpora.ru/new/sbornik2005/04reznikova.pdf> (Дата обращения: 01.06.2021).