6 – TOM 5 – SON / 2023 - YIL / 15 - MAY

СОЗДАНИЕ И ЗНАЧЕНИЕ ЯЗЫКАЛИНГВИСТЕЧЕСКОГО КОРПУСА

Гули Тоирова Ибрагимовна
доктор филологических наук, доцент, Бухарский государственный университет, e-mail: tugulijon@mail.ru
Виркани Музаффар Абдуллло
Таджикского национального университета, старший преподаватель, e-mail: virkan76@gmail.com

Аннатация. *В статье рассматривается трансформация языка в язык Интернета, компьютерные технологии, математическая лингвистика, ее продолжение и становление и развитие компьютерной лингвистики, в частности вопрос моделирования естественных языков для искусственного интеллекта. В частности, исследуется вопрос лингвистического и экстралингвистического разделения специальных тегов для маркировки текстов и их компонентов. Определены требования к кодированию важной текстовой информации. В статье рассматривается основное назначение корпуса как сложного лингвистического источника, а также тот факт, что он в основном содержит два вида информации и ее типы. Национальный корпус, образовательный корпус и параллельный корпус обсуждаются в рамках предмета компьютерной лингвистики. Было подчеркнуто, что их лингвистическая и экстралингвистическая маркировка, разработка алгоритмов формирования корпусов и создание корпусной лингвистической поддержки являются общественной потребностью.*

Ключевые слова: *корпус, искусственный интеллект, лексическая информация, морфологический признак, слово.*

Annotation. *The article discusses the transformation of language into the language of the Internet, computer technology, mathematical linguistics, its continuation and the formation and development of computational linguistics, in particular, the issue of modeling natural languages for artificial intelligence. In particular, the issue of linguistic and extralinguistic separation of special tags for marking texts and their components is investigated. The requirements for coding important textual information are determined. The article discusses the main purpose of the corpus as a complex linguistic source, as well as the fact that it mainly contains two types of information and its types. National corpus, educational corpus and parallel corpus are discussed under the subject of computational linguistics. It was emphasized that their linguistic and extralinguistic marking, the development of corpus formation algorithms and the creation of corpus linguistic support are a social need.*

Key words: *corpus, artificial intelligence, lexical information, morphological feature, word.*

1. Введение. Благодаря современным информационным технологиям искусственный интеллект дал широкий спектр преимуществ в использовании языка. Он способен делать множество вещей, на которые способен человеческий интеллект. Электронные источники, созданные искусственным интеллектом, предназначены для обеспечения безопасности людей и снижения их веса. Среди наиболее актуальных проблем - перевод узбекского языка на Интернет и электронный язык, а также расширение электронных ресурсов на национальных языках (корпус узбекских языков, электронные словари, содержание веб-сайтов).

Вопрос(ы) исследования

Ранее мы упоминали, что языки, достигшие мировой языковой цивилизации, уже проделали работу по обработке информации с использованием компьютерных технологий, машинного перевода, электронной лексикографии, созданию тезаурусов, созданию языковых корпусов. Английский, русский, арабский, французский, немецкий, испанский и таджикский — лишь некоторые из них. Также установлены научно-теоретические аспекты создания языкового корпуса в системе Интернет на этих языках, подчеркнута необходимость активизации усилий по превращению узбекского языка в «понятный» Интернету.

В мировой лингвистике создание языковых корпусов в Интернете является основным средством сохранения того или иного языка ко второму десятилетию двадцать первого века, расширения сферы его исследования и демонстрации языковых навыков. В частности, компьютерная технология, которая является великим изобретением двадцатого века, открывает дверь в широкий спектр возможностей для лингвистики, а также других областей и накладывает огромные задачи на компьютерный язык, появление компьютерной лингвистики имеет решающее значение для Успех естественных языков.

В глобальных лингвистических исследованиях изучение лингвистического моделирования языка, разработка алгоритмов лемминга слов и тегов, а также использование в электронной форме устных и письменных памятников, образцов духовного наследия, созданных на конкретном языке, с целью повышения использование национального и культурного наследия. Особое внимание уделяется обработке информации с помощью компьютерных технологий, разработке необходимого программного и методического обеспечения для внедрения информационных ресурсов, развитию языкового корпуса в сети Интернет и на этой основе научно-теоретическим аспектам национального языка. корпус.

2. Обзор литературы

Корпус является предметом корпусной лингвистики. Этот термин по-разному определяется в научной литературе. Например, в английском языке он используется с такими терминами, как лингвистический корпус или текстовый корпус. Признание научных исследований А.Н. Хомский, Г.Н.Луч, Ч.Ф.Меер, Дж.Синклер, М.З.Курди в решении таких проблем, как создание национального корпуса определенного языка,

его аналитическая технология, развитие области корпусной лингвистики должны (Мохамед Закария Курди, 2016 г. ; Тойрова, 2020, с.57; Чарлез, 2004, с.7; Шомску, 1962).

Джон Синклер определяет термин «корпус» следующим образом: «Корпус состоит из фрагментов текстов в электронной форме, отобранных по видимым критериям для изучения языка или языкового разнообразия, которые должны быть представлены в качестве источника информации» (Sinclair, 2004).).

Большой массив массивных текстов русской корпусной лингвистики, принципы формирования корпусов, лингвистическая база данных В.Г. Бритвин, В.П. Захаров, И.А. Мельчук, А.Б. Кутузов, Р.Г. Котов, Л.И. Беляева , Отраженные в целевых исследованиях Е.В.Недошивиной, В.В.Рыкова, В.Плунгяна (Бритвин, 1983; Блумфилд, 1968; Беляева, Чижаковский, 1983; Захаров, 2011; Недошивина, 2006; Рыков, 2005; Плунгян, 2005; Кутузов, 2017; Котов, 1977).

Русский ученый В.П. Захаров объясняет термин «корпус» следующим образом: «корпус — совокупность лингвистических данных единиц языка, составленных на основе устных и письменных текстов» (Захаров, 2011).

Х.Исхакова, С.Мухаммедов, С.Риза о лингвостатистическом анализе текста в узбекском языкознании, лексикографической обработке, лингвистическом обеспечении программы автоматического редактирования, лингвистических модулях программы редактирования и анализа, синонимической лексике национального корпуса , лингвистические основы авторского корпуса. С.Мухаммедова, Б.Менглиев, Д.Уринбаева, А.Пулатов, У.Дысимова, Г.Валиева, Г.Джуманазарова, Н.Абдурахмонова, Ш.Хамроева, М.Абжалова, А.Эшмоминов, О.Холиёров, Р. Работа Каримова заслуживает внимания. Наши ученые, такие как С.Каримов, С.Мухаммедова, Ш.Хамроева, проводили исследования по специальности 10.00.01 корпусной лингвистики.

Узбекские лингвисты определяют термин «корпус» следующим образом: узбекские лингвисты интерпретируют термин «корпус» следующим образом: «корпус – это совокупность языковых единиц, составляющих совокупность текстов, собранных с определенной целью» (Эшмуминов, 2019), а набор письменных или устных текстов, хранящихся в электронном виде на языке, помещенный в компьютеризированную поисковую систему» (Бонгерс, 1947). Исследования в области узбекского языкознания так описывают сущность корпуса: «Корпус – это способность представить существующую информацию в виде текста; возможность предоставить максимум информации в зависимости от размера кейса; это возможность многократно использовать данные однажды созданного корпуса для решения различных задач» (Пулатов, 2011).

«Корпус – это совокупность текстов, подлежащих поисковой системе с целью определения характеристик языковых единиц, письменных или устных, хранящихся в электронном виде на естественном языке, размещенных в компьютерной поисковой

системе с программным обеспечением на основе -линейная или автономная система» (Менглиев, Бобожонов и Хамроева, 2018) источник.

О.Халиёров, проводивший исследование «образовательного корпуса», в своей работе констатирует следующее: Учебный корпус узбекского языка – это корпус, предназначенный для обучения возможностям узбекского языка, имеет лингводидактический характер, содержит электронные тексты, выступает в качестве специального сайта» (Хамроева, 2018).

О параллельном корпусе Р. Каримов говорит: «параллельный электронный аналог переводных текстов; состоит из нескольких «оригинальных текстов и одного/нескольких их переводов» (Каримов, 2021).

Языковые корпуса можно разделить на разные формы по структуре, назначению, устойчивости, вариативности. Например, В.П. Захаров перечисляет следующие формы: «по форме хранения данных: звуковая, письменная, смешанная; по языку текста: одноязычные и многоязычные; по жанру: литературный, диалектный, устный, публицистический, смешанный; по подъезду к зданию: свободный, хозяйственные постройки, закрытый; по назначению: исследовательская, иллюстративная; по изменчивости: динамичные и стабильные; маркированные и немаркированные по признаку наличия дополнительной информации (аннотированные)» (Захаров, 2011).

В.В. Рыков же акцентирует внимание на следующих аспектах классификации корпусных типов: «По уровню и структуре данных, по хронологическому признаку (позиции) языка, по языку употребления, по до цели использования» (Рыков, 2005).

В своем запрете Ш. Хамроева делит корпус на следующие типы: «По определенному периоду языка или определенному типу его возникновения (жанр, стиль, социальная или возрастная группа, язык писателя или ученого); по типу языкового знака; по типу речи: письменная, устная, смешанная; они выглядят как мультимодальный корпус, корпус специальных текстов» (Хамроева, 2018).

«Специализированный корпус: группа текстов определенного типа: газетный текст, научные статьи; общий корпус; сравнительный корпус; параллельный корпус; учебный корпус; дидактический корпус», — говорит У.Холиёров (Холиёров, 2021).

3. Методология

Каждый академик определял лингвистические корпуса со своей точки зрения и классифицировал их по-разному. Какие черты узбекского лингвистического корпуса нашли отражение в нем и какие корпуса сейчас создаются?

Создание национального корпуса узбекского языка является относительно новым направлением как в узбекской лингвистике, так и в современных информационных технологиях. Языковой корпус является основным источником и мощным информационным ресурсом для составления больших словарей. Языковой корпус позволяет быстро создавать и обрабатывать словари с помощью компьютера. Важность корпуса в области лексикографии заключается в том, что ни один инструмент не может сравниться с корпусом в определении периода и частоты

использования слова. В ближайшем будущем потребность в словаре сегодня для студента, изучающего язык, или исследователя, изучающего какой-либо его аспект, несомненно, переместится на корпус.

4. Результаты и обсуждение

Лингвисты Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои сейчас работают над научным и практическим проектом под названием «Образовательный корпус». Создание образовательного корпуса по узбекскому языку направлено на поэтапное формирование данных на основе зарубежного опыта и включает в себя электронный учебник, содержащий современную лексику узбекского литературного языка, многоязычных носителей и непереведенных лексических единиц узбекского языка, а также комплект мультимедийных продуктов, включающий аудио- и видеоматериалы, а также мобильное приложение, направленные на формирование навыков правильного произношения на узбекском языке.

Система образования позволяет учащимся углубленно изучать узбекский язык как государственный, второй язык и иностранный язык. Пользователи могут свободно изучать узбекский язык благодаря электронному материалу учебного корпуса, который включает в себя аудио-, видео-, мультимедийные приложения, программы для произношения и правописания, а также словари для электронного обучения. В отличие от других учебных программ, этот комплекс направлен на развитие способности использовать узбекский язык в необычных контекстах. Новички, студенты, родители, преподаватели и изучающие узбекский язык могут извлечь из этого пользу. Это будет способствовать формированию научно-технических ресурсов, которые обеспечат экономический рост и социальное развитие республики.

В результате направленности научных исследований в области узбекской компьютерной лингвистики на обработку узбекского языка с использованием современных информационных технологий появляются первые появления национального корпуса в виде практической работы.

Ниже представляем план формирования Центрального банка для создания национального корпуса узбекского языка:

МБ в Microsoft Access Есть два способа создать таблицу МБ в MS Access MBBT. Самый простой способ создать МБ — создать все необходимые таблицы, формы и отчеты с помощью Мастера МБ. Однако вы можете создать пустую ББ и затем добавить в нее таблицы, формы, отчеты и другие объекты — это наиболее удобный способ, но он требует отдельного определенного объекта ББ. В обоих случаях есть возможность модифицировать и расширять созданный МБ. Чтобы создать новую МБ, выберите «Создать», «Новая база данных», а затем «Создать» в меню «Файл» (рис. 1 (а)).

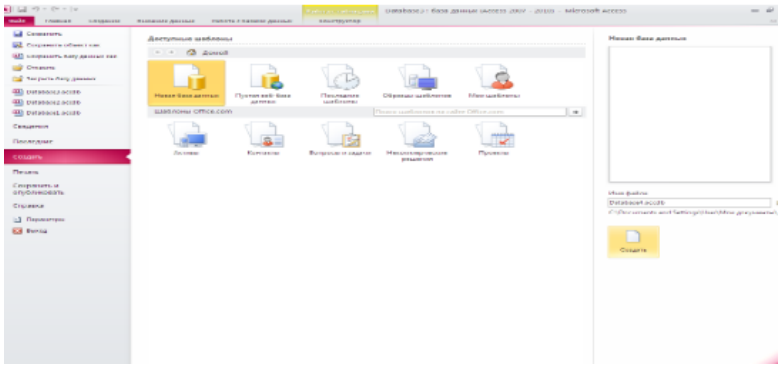

Рис. 1. (а). окно МБ

На экране появится окно «Создать МБ». В поле Имя введите названия полей и выберите соответствующие типы, затем сохраните таблицу и переключитесь в табличный режим (Рисунок 1 (а)).

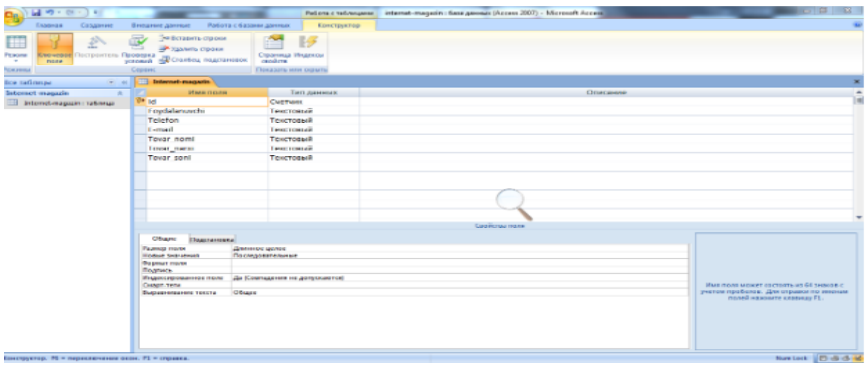

Рисунок 2 (а). окно МБ

Затем в появившемся окне вводим табличные значения, обращая внимание на тип поля. Следует вводить только соответствующие типы данных. Например, <force> нужно сделать в одном столбце, во втором столбце; нойлой. Человек обязан сделать что-либо против своей воли, по принуждению или по необходимости» (рис. 2 (б)).

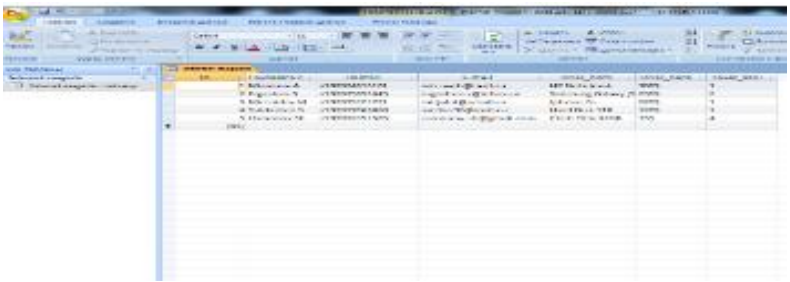

Рисунок 3 (б). окно МБ

Access позволяет редактировать поля и записи таблицы. Вы также можете изменять, добавлять и удалять поля в режиме мастера и таблицы. Ввод и редактирование данных в таблице производится в табличном режиме. Доступ имеет следующие типы данных: Основной тип, Число, Данные/время, Да/Сеть и Быстрый запуск. При создании МБ необходимо обращать внимание на тип данных и при вводе указывать соответствующие данные (рисунок 4).

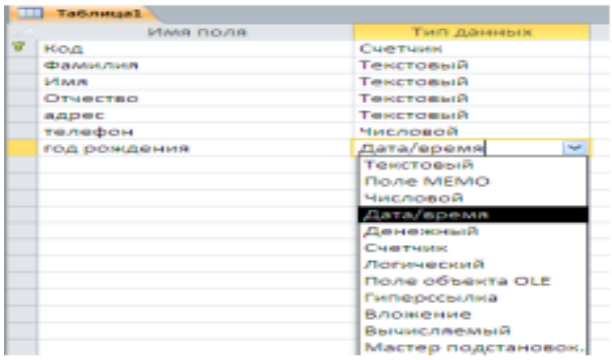

Рисунок 4 (а). Окно ввода и редактирования данных

Чтобы добавить новую запись, сначала открывается таблица или форма. Новые данные вводятся в последней строке. Например, [ф], [статус. ф.]. Редактирование данных осуществляется подобно простой таблице (рис. 5).

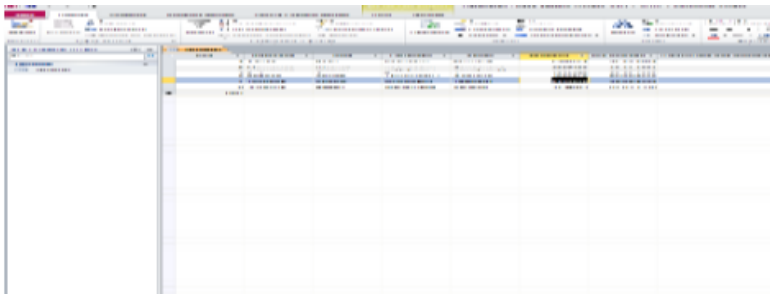

Рисунок 5 (б). Окно ввода и редактирования данных

Чтобы удалить записи, выберите соответствующую запись, щелкните правой кнопкой мыши и выберите Удалить запись.

МБ хранит миллионы записей, из которых в любой момент можно найти нужную информацию. Данные в таблицах МБ должны иметь простые средства поиска необходимой информации. Поиск и сортировка осуществляется в табличном режиме и по специальным запросам. Создается соответствующий запрос, результатом которого являются требуемые записи.

Поиск информации осуществляется через запросы, и в результате запроса мы имеем новую таблицу, удовлетворяющую заданным условиям (рисунок 6).

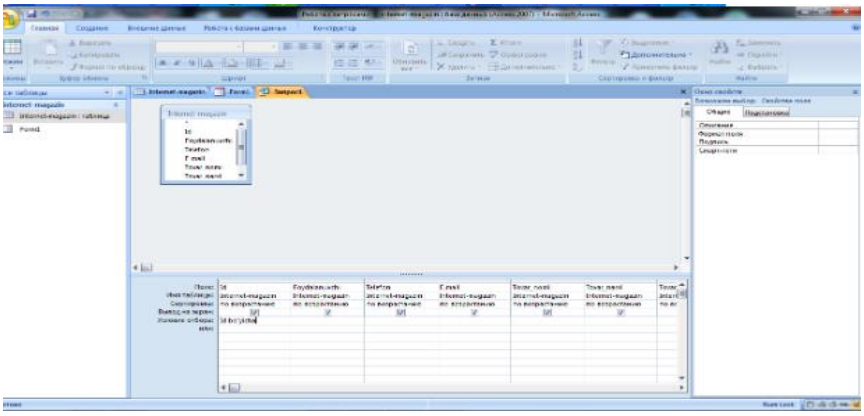

Рисунок 6. Окно запроса поиска информации

В МБ информацию можно сортировать, слова располагать по алфавиту или нумеровать. Сортировка сделана для удобства поиска данных. Обычно таблица сортируется по значению ключевого поля. Сортировка может производиться по одному или нескольким полям. Для этого выберите необходимые поля и выберите условие сортировки. Моделирование базы данных выполняется шаг за шагом (Fries, 1969).

Процесс экспертизы включает в себя сбор материалов, их оформление в виде технического задания. Они обосновывают целесообразность создания банка и базы данных. Следующие факторы были определены и названы ключевыми факторами:

- часто используемая информация;
- предоставление пользователям интерактивного доступа к данным;
- наличие сложных связей между данными;
- необходимость обновления системы.

Материалы, содержащие выводы и предложения по созданию банка и базы данных исходя из определенных условий и возможностей, включаются в ТЭО проекта, а также являются основанием для формирования технических условий на разработку системы баз данных.

Например, арабская лингвистика уже включила арабский текст в лингвистическую базу данных, предназначенную для подключения к основным поисковым системам и программам поиска данных, чтобы облегчить анализ документов, написанных на арабском языке. Однако он не подходит для стандартных автоматизированных методов анализа, учитывающих арабскую письменность, на которой традиционно говорят. В такой языковой базе арабские слова часто содержат грамматические элементы, обозначающие такие признаки, как направление глагола, дополнение, лицо, число, род и т. д. Эта система автоматически выполняет следующие действия: 1) формирует языковую форму слова; 2) определяет части речи; 3) нормализует орфографические нормы, в том числе удаление гласных и имен существительных, унификацию форм хамзы и постоянно ломаных неправильных «разорванных» форм множественного числа; 4) может работать на персидском (персидском и дари), пушту и урду (рис. 7).

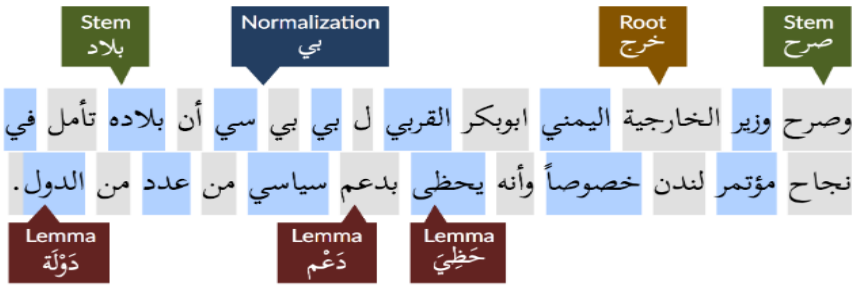

Рисунок 7. Окно автоматического действия

5. Вывод

Короче говоря, корпусная лингвистика — это самая передовая ветвь лингвистики, а корпус — необходимый инструмент для лингвистов; устные, письменные памятники являются источником информации, отражающей национально-культурное наследие. Корпус представляет собой совокупность текстов, подлежащих поисковой программе, а четко определенный корпус служит устойчивой лингвистической базой в обеспечении эффективности лингвистического исследования. Являясь продуктом искусственного интеллекта, лингвистический корпус включает в себя электронный словарь, портал переводов, терминологическую базу, виртуальную (электронную) библиотеку, электронное правительство, электронные издательства, электронные учебники и учебные пособия. Общий вид узбекского национального корпуса разделен на несколько окон и правую и левую колонки. В нем будут следующие окна: «Лексический поиск», «Морфологический поиск», «Синтаксический поиск». Слова и фразы из него автоматически анализируются за считанные секунды. Лингвистические и экстралингвистические маркировки создаются в едином формате выражения данных в узбекском национальном корпусе, а также в мировых языковых корпорациях. Будет проведен пересмотр теоретических основ морфологической и синтаксической разметки на основе академической грамматики, практическая работа, связанная с сокращением системы семантических тегов разметки. Значение гнезда в корпусе несравнимо, ведь от гнезда корпуса зависит ширина или узость доступа к корпусу. Идеальная планировка – залог широкого выбора вариантов, универсальности жилья.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА:

1. Belyaeva, L.I., Chizhakovsky, V.A., (1983). Thesaurus in automatic text processing systems., Chisinau.
2. Bloomfield, L., (1968). Language. Moscow, "Progress".
3. Bongers, H., (1947). The history and principles of Vocabulary control, Woerden: WOCOPI.
4. Britvin, V.G., (1983). Applied modeling of syntagmatic semantics of scientific and technical text (by the example of automatic indexing), Moscow State University.
5. Charlez, Meyer, (2004). English corpus linguistics: An introduction. Cambridge University Press, UK, 168 p.
6. Eshmuminov, A., (2019). Synonymous database of the Uzbek language national corpus. Dissertation of PhD in Philology, Tashkent.
7. Fries, Ch.C., (1969). The structure of English. An introduction to the construction of English sentences, London.

8. Hamroeva, Sh., (2018). Linguistic bases of creation of the author's corpus of the Uzbek language: Author's Abstract of the Dissertation of PhD in Philology, Tashkent.
9. Karimov, R., (2021). Linguistics and programming issues of creating a parallel corpus of Uzbek and English, Author's Abstract of dissertation of PhD, Bukhoro, 151 p.
10. Kholiyorov, O., (2021). Linguistic bases of formation of educational corpus of Uzbek language. Author's Abstract of the Dissertation of PhD in Philology, Termiz.
11. Kotov, R.G., (1977). Linguistic aspects of automated control systems. Moscow, Nauka.
12. Kutuzov, A.B., (2017). Corpus linguistics. Retrieved from: http://lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf ;
13. Leech, G., (1991). The State of Art in Corpus Linguistics, English Corpus Linguistics, London.
14. Melchuk, I.A., (1985). Word order in the automatic synthesis of the Russian word (preliminary messages), Scientific and technical information, 12:12-36.
15. Mengliev, B., (2018). Is the Uzbek language corpus being created? Ma'rifat newspaper. April 3, 2018, retrieved from: http://marifat.uz/marifat/v_pomosh_uchitelu-marifat/savol/1142.htm.
16. Mengliev, B., Bobojonov, S., Hamroeva Sh., (2018). Uzbek National Corpus. April 26, 2018, retrieved from: http://marifat.uz/marifat/ruknlar/fan/1241.htm.
17. Mohamed Zakaria Kurdi, (2016). Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax, Great Britain, USA: Wiley-ISTE, 300 p.
18. Nedoshivina, E.V., (2006). Programs for working with text corpuses: an overview of the main corpus managers. Study guide, St. Petersburg, 26 p.
19. Plungyan, V., (2005). Why are we making the National Corpus of the Russian language? [Electronic resource], Notes of the Fatherland, 2:20, retrieved from: http://magazines.russ.ru/oz/2005/2/2005_2_20-pr.html.
20. Pulatov, A. Q., (2011). Computer Linguistics. Tashkent, Akademnashr, 520 p.
21. Rykov, V.V., (2005). A course of lectures on corpus linguistics. URL: http://rykov-cl.narod.ru/c.html.
22. Shomsku, N., (1962). The logical basis for linguistic theory, Proceedings of the IX International Congress of Linguists.
23. Sinclair, D., (2004). How to use corpuses in teaching a foreign language, Preface to the book, Studies in Corpus Linguistics, 12, VIII, 308 pp. retrieved from: http://www/ruscorpora.ru/corpora-info.html.

24. Toirova, G., (2019). The Role of Setting in Linguistic Modeling. International Multilingual Journal of Science and Technology, 4(9):722-723, available at: http://imjst.org/index.php/vol-4-issue-9-september-2019/.

25. Toirova, G., (2020). About the technological process of creating a national corpus. Foreign languages in Uzbekistan, 2(31):57-64, available at: https://journal.fledu.uz/uz/2-31-2020.

26. Toirova, G., (2020). The importance of the interface in the creation of the corpus. Ínternauka, 7, DOI: https://doi.org/10.25313/2520-2057-2020-7-5944.

27. Toirova G.,Yuldasheva M., Elibaeva l. Importance of Interface in Creating Corpus. // International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2S10, September 2019. –P.352-355. (scopus)

28. Toirova G., Jurayeva O., Abulova Z., Norova M., Norova F. Application of Innovative Technologies in Teaching Process. // International Journal of Psychosocial Rehabilitation, Vol. 24, Special Issue 1, 2020. ISSN: 1475-7192.–P.386-390. (scopus)

29. Zakharov, V.P., (2011). Corpus linguistics: a textbook for students of humanitarian universities, Irkutsk, 161 p.