O‘ZBEKISTON RESPUBLIKASI
RAQAMLI TEXNOLOGIYALAR VAZIRLIGI
OLIY TA’LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI
SAMARQAND FILIALI

“O‘ZBEK TILINING MILLIY KORPUSI: MUAMMOLAR VA VAZIFALAR”

mavzusidagi xalqaro ilmiy-amaliy konferensiya

(Samarqand shahri, 2023 yil 16-17 mart)

AXBOROT XATI

SAMARQAND – 2023

AXBOROT XATI

O‘zbekiston Respublikasi Prezidentining 2019 yil 21 oktyabrdagi “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi PF-5850-son Farmoni, 2020 yil 6-oktyabrdagi PQ-4851-sonli “Axborot texnologiyalari sohasida ta’lim tizimini yanada takomillashtirish, ilmiy tadqiqotlarni rivojlantirish va ularni IT-industriya bilan integrasiya qilish chora-tadbirlari to‘g‘risida”gi qarori hamda axborot texnologiyalari sohasini rivojlantirishning ustuvor vazifalarini amalga oshirish maqsadida, **2023 yil 16-17 mart** kunlari Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filialida **“O‘zbek tilining milliy korpusi: muammolar va vazifalar”** mavzusida xalqaro ilmiy-amaliy konferensiya o‘tkaziladi.

Konferensiya O‘zbekiston Respublikasi Vazirlar Mahkamasi huzuridagi O‘zbek tilini rivojalantirish jamg‘armasi tomonidan moliyalashtirilgan “O‘zbek tilining milliy korpusini loyihalash va dasturiy majmua ishlab chiqish” mavzusidagi amaliy loyiha doirasida tashkil etiladi.

Konferensiya tashkilotchilari: TATU Samarqand filiali dasturiy injiniring va SamDU o‘zbek tilshunosligi kafedralari.

***Dasturiy qo‘mita raisi*:** filologiya fanlari doktori, professor
Suyun Amirovich Karimov

Koordinator: Muhammadsolih Tursunov, +99897 395 57 29

KONFERENSIYA QUYIDAGI SHO‘BALAR BO‘YICHA FAOLIYAT YURITADI:

- **1-Sho‘ba:** O‘zbek tilining milliy korpusi: natijalar, muammolar, vazifalar.
- **2-Sho‘ba:** Korpus tilshunosligining nazariy va amaliy masalalari.
- **3-Sho‘ba:**Tilshunoslikda raqamli va axborot texnologiyalar.

Konferensiyaga yosh olimlar, mustaqil izlanuvchilar, tayanch doktorantlar, doktorantlar, ilmiy-tadqiqot institutlari xodimlari hamda oliy ta’lim muassasalari professor-o‘qituvchilari taklif etiladi.

Konferensiyada bevosita (konferensiya zalida) hamda onlayn tarzda (videokonferensiyada) ishtirok etish mumkin.

KONFERENSIYADA ISHTIROK ETISH UCHUN:

Ro‘yxatga olish kartasi (Ilova 1) va maqola materiallari alohida faylda qo‘yilgan talablar asosida (Ilova 2) elektron shaklda *unc_conf@samtuit.uz* yoki *muhammadsolih927@gmail.com* elektron manziliga **2023 yilning 5 martiga qadar** yuboriladi.

Maqolaning elektron nusxasi quyidagi shaklda nomlanadi: sho‘ba raqami, birinchi muallifning familiyasi, masalan: ***1_tursunov.docx***

Maqolalarga qo‘yiladigan talablar:

Hajmi to‘liq 5 betdan kam bo‘lmasligi lozim. Matn 1 intervalda (chapdan 3 sm, yuqoridan 2 sm, quyidan 2 sm, o‘ng tomondan 1,5 sm, varaq bichimi A4 formatda, 210x297mm), **Microsoft Word** (*.**docx**) muharririda, matnlar **Times New Roman**, **14** o‘lchamli shriftda, chizma yoki rasmlar varaqning o‘rtasida joylashtiriladi, chizma yoki rasmlarning tagida izohlari 12 pt o‘lchamida varaqning o‘rtasidan yoziladi; maqola nomi 12 so‘zdan ortmasligi, bosh harflarda qoraytirilib, varaqning o‘rtasiga yoziladi; maqola nomi maqola shakllantirilgan til va ingliz tiliga tarjima qilingan holatda berilishi; maqola sarlavhasidan 1 interval pastda muallifning familiyasi, ismi, sharifi, ularning cheti – chap tomonida (*) havolasi ostida muallifning ilmiy darajasi, unvoni, tashkilot nomi va elektron manzili ko‘rsatilishi lozim. 1 interval tashlanib maqola annotatsiyasi, kamida 5-7 ta kalit so‘z maqola tili va ingliz tilida keltirilishi, maqola matni kalit so‘zlardan keyin 1 interval pastdan berilishi kerak (Ilova 2).

Maqolalar **o‘zbek**, **rus** yoki **ingliz** tillarida qabul qilinadi.

Maqola matni ilmiy, texnik, grammatik va stilistik tahrir qilingan bo‘lishi shart. Maqoladagi ma’lumot, fakt va statistik ko‘rsatkichlarning to‘g‘riligiga mualliflar mas’ul. Maqolada albatta jadval (chizma yoki rasm) manbalari aniq ko‘rsatilishi, qisqartma so‘zlarga izoh berilishi lozim. Maqola ichidagi havolalar “[1]” kabi tartibda belgilanadi. Maqola so‘ngida foydalanilgan adabiyotlar foydalanish ketma-ketligi bo‘yicha yozilishi kerak.

Yuqoridagi talablarga javob bermaydigan, o‘z vaqtida topshirilmagan, kamchiliklari mavjud bo‘lgan ilmiy maqolalar to‘plamga kiritilmaydi. Tashkiliy qo‘mita maqola matnini qisqartirish, qisman tuzatish kiritish, sho‘balarga joylashtirish huquqiga ega.

Maqolalar to‘plami konferensiya boshlanish kuniga qadar electron tarzda chop etiladi.

Ilova 1.

Ishtirokchini ro‘yxatga olish kartasi

1. Familiyasi, Ismi, Sharifi _____
2. Ilmiy darajasi _____
3. Ilmiy unvoni _____
4. Lavozimi _____
5. Tashkilot nomi (to‘liq va qisqartirilgan) _____
6. Elektron pochta _____
7. Telefon raqami _____
8. Davlat _____
9. Shahar _____
10. Mualliflar F.I.Sh. va maqolaning nomi _____
11. Oflayn shaklda ishtirok etuvchi to‘g‘risida ma’lumot _____
12. Sirtdan ishtirok etuvchi to‘g‘risida ma’lumot _____

Ilova 2.

O‘ZBEK TILI MILLIY KORPUSI UCHUN MATNLARNI FORMATLASH
FORMATTING TEXTS FOR THE NATIONAL CORPS OF THE UZBEK LANGUAGE

*Tursunov Muhammadsolih Sa’din o’g’li
*Muhammad al Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filiali, Samarqand, O‘zbekiston
muhammadsolih927@gmail.com

Annotatsiya. Ushbu maqolada o‘zbek tili milliy korpusiga matnlarni kiritishda foydalanilgan usullarni tavsiflash va kodlashga umumiy yondashuv muhokama qilinadi. Umumiy format mavjud matn formatlarining xilma-xilligi va nomuvofiqligi bilan asoslanishi mumkin. Korpusda matnlarni saqlash uchun JSON formatdan foydalanish orqali korpus qidiruv tezligini oshirish va kengayuvchanlikdagi nazariy va texnik muammolarni bartaraf etish mumkin. Korpusga Alpomish dostoning matnlari kiritilishi tavsiflangan.

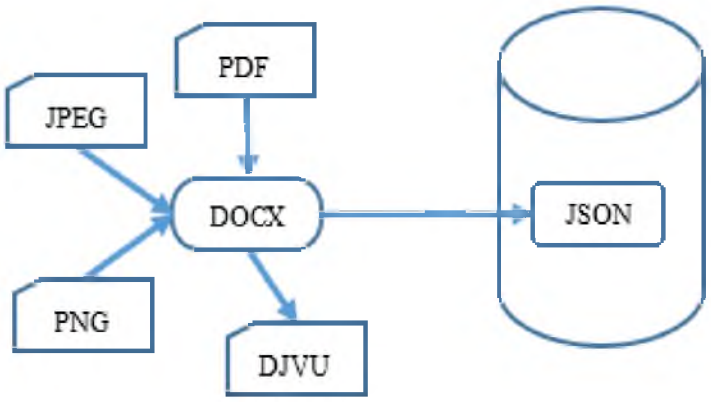
Abstract. This article discusses the general approach to the description and coding of the methods used in the inclusion of texts in the national corpus of the Uzbek language. A common format can be justified by the diversity and incompatibility of existing text formats. By using the JSON format to store texts in the corpus, it is possible to increase corpus search speed and overcome theoretical and technical problems of scalability. The inclusion of the texts of the Alpomish epic into the corpus is described.

Kalit so‘zlar: *korpus, formatlash, fayl, matn, Alpomish dostoni, token, razmetka, teg, tegger, JSON format, DOCX format.*

Keywords: *Corpus, formatting, file, text, Alpomish epic, token, markup, tag, tagger, JSON format, DOCX format.*

Bugungi kunda korpuslar lug‘atlar va grammatika kabi tilshunoslikning ajralmas qismiga aylandi. Korpus paydo bo‘lganidan so‘ng tilshunoslik fanlari o‘zgarib ketdi, aytish mumkinki, butun tilshunoslik korpus tilshunosliliga aylandi. Eng taniqli va tan olingan lingvistik korpuslarga namuna sifatida quyidagilarni keltirish mumkin: Rus milliy korpusi (https://ruscorpora.ru/new/), Britaniya milliy korpusi (http://www.natcorp.ox.ac.uk/, https://www.english-corpora.org/bnc/), Turk milliy korpusi (https://www.tnc.org.tr/), Amerika milliy korpusi (http://www.anc.org/) va boshqalar[1].

Matnlar turli xil PDF, rasmli, dokumentli va boshqa formatlarda bo‘ladi. Korpusga matnlarni kiritishdan avval, mavjud matnli fayllarni Microsoft Office ning 2010 yil va undan yuqori bo‘lgan versiyasidagi *.docx formatiga o‘tkazish kerak bo‘ladi. Boshqa formatdagi matnlarni *.docx formatiga maxsus dasturlar yordamida o‘tkaziladi va *.docx formatiga o‘tkazish jarayonida matnning asl holati buzilishi mumkin. Bunda matndagi imloviy xatolar qo‘l mehnati yordamida matnning asl holati bilan bir xillikka keltiriladi. Undan so‘ng matnni korpusga yuklash mumkin bo‘ladi. Ushbu tadqiqotda korpusda matnlarni saqlash uchun JSON formatdan foydalanilgan (1-rasm).



1-rasm. *Matnlar formati va korpusga saqlash formati*

Foydalanilgan adabiyotlar ro’yxati

1. A.B.Karshiev, S.A.Karimov, M.S.Tursunov, Development of a Modern Corpus of Computational Linguistics // Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT), DOI: 10.1109/ICISCT50599.2020.9351376, 2021.

МИНИСТЕРСТВО ЦИФРОВЫХ ТЕХНОЛОГИЙ
МИНИСТЕРСТВО ВЫСШЕГО ОБРАЗОВАНИЯ, НАУКИ И ИННОВАЦИЙ
РЕСПУБЛИКИ УЗБЕКИСТАН

САМАРКАНДСКИЙ ФИЛИАЛ ТАШКЕНТСКОГО УНИВЕРСИТЕТА
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ИМЕНИ
МУХАММАДА АЛ-ХОРАЗМИЙ

ИНФОРМАЦИОННОЕ ПИСЬМО

о международной научно-практической конференции

“НАЦИОНАЛЬНЫЙ КОРПУС УЗБЕКСКОГО ЯЗЫКА: ПРОБЛЕМЫ И ЗАДАЧИ”

(г. Самарканд, 16-17 марта 2023 года)

САМАРКАНД – 2023

ИНФОРМАЦИОННОЕ ПИСЬМО

В целях выполнения Постановления Президента Республики Узбекистан ПП-5850 от 21 октября 2019 года «О мерах по повышению влияния и позиции узбекского языка как государственного языка», ПП-4851 от 6 октября 2020 года «О мерах по дальнейшему совершенствованию системы образования в области информационных технологий, развитию и интеграции научных исследований с IT-индустрией» и в целях реализации приоритетных задач развития сферы информационных технологий **16-17 марта 2023 года** в Самаркандском филиале Ташкентского университета информационных технологий имени Мухаммада ал-Хоразмий состоится международная научно-практическая конференция на тему "**Национальный корпус узбекского языка: проблемы и задачи**".

Конференция организуется в рамках проекта прикладных исследований «Проектирование национального корпуса узбекского языка и разработка программного комплекса», финансируемого Фондом развития узбекского языка при Кабинете Министров Республики Узбекистан.

Организаторы конференции: кафедры программной инженерии Самаркандского филиала ТУИТ и узбекского языкознания СамГУ.

Председатель программного комитета: доктор филологических наук, профессор Суюн Амирович Каримов

Координатор: Мухаммадсолих Турсунов, +99897 395 57 29

СЕКЦИИ КОНФЕРЕНЦИИ:

- **Секция 1:** Национальный корпус узбекского языка: результаты, проблемы, задачи;
- **Секция 2:** Теоретические и прикладные вопросы корпусной лингвистики;
- **Секция 3:** Цифровые и информационные технологии в лингвистике.

На конференцию приглашаются молодые ученые, соискатели, докторанты, сотрудники научно-исследовательских институтов, а также преподаватели высших учебных заведений.

В конференции предусмотрено очное (в конференц-зале) и онлайн (видеоконференция) участие.

ДЛЯ УЧАСТИЯ В КОНФЕРЕНЦИИ НЕОБХОДИМО:

Отправить электронном виде на адрес электронной почты *unc_conf@samtuit.uz* или *muhammadsolih927@gmail.com* регистрационную карту (см. Приложение 1) и материалы статьи, оформленной строго в соответствии с требованиями (см. Приложение 2) в срок до **5 марта 2023 года**.

Электронный вариант статьи следует представить в отдельном файле с указанием номера секции и фамилии первого автора, например: ***1_tursunov.docx***.

Требования к статьям:

Размер статьи и тезисов должен составлять не более 5 страниц, интервал 1. Границы страницы (слева 3 см, сверху 2 см, снизу 2 см, справа 1,5 см, размер листа в формате А4, 210х297 мм), в редакторе **Microsoft Word (*.docx)**, текст шрифтом **Times New Roman**, **14-пт.** Рисунки или изображения размещаются по середине, подписи к рисункам пишутся внизу рисунков по середине размером шрифта 12 пт. Название статьи должна быть написана заглавными буквами и написана в середине страницы и не превышать 12 слов, название статьи указать на языке написанном в оригинале и перевод на английском языке в следующих строках указать фамилию, имя, отчество автора – с левой стороны указать ссылку(*) и в следующий строке указать научную степень и звание, полное название организации, адрес электронной почты. Со следующей строки привести аннотацию статьи и ключевые слова не менее 5-7 слов на языке оригинале и на английских языках (см. Приложение 2).

Статьи принимаются на **узбекском, русском** или **английском** языках.

Текст статьи должен быть научно, технически, грамматически и стилистически отредактирован. Авторы несут ответственность за достоверность информации, фактов и статистики в статье. В статье должен быть четко указан источник таблицы (рисунок или картинка), сокращения должны быть расшифрованы. Ссылки внутри статьи помечаются в том же порядке, что и «[1]». Ссылки, используемые в конце статьи, должны быть написаны в порядке их использования.

Научные статьи, не соответствующие вышеуказанным требованиям, не представленные вовремя и имеющие недостатки, не будут включены в сборник. Оргкомитет имеет право сокращать тексты статей, вносить частичные исправления, размещать по секциям.

Сборник статей будет опубликован в электронном виде накануне дня конференции.

Приложение 1.

Регистрационная форма участника

1. Фамилия, Имя, Отчество _____
2. Ученая степень _____
3. Ученое звание _____
4. Должность _____
5. Организация (полное и сокращенное название) _____
6. Электронная почта _____
7. Контактный телефон _____
8. Страна _____
9. Город _____
10. Авторы и название статьи _____
11. Участие в офлайн формате с докладом _____
12. Участие в онлайн формате _____

Приложение 2.

ПРИМЕР

ОФОРМЛЕНИЕ ТЕКСТОВ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА
FORMATTING TEXTS FOR THE NATIONAL CORPS OF THE UZBEK LANGUAGE

[*]Турсунов Мухаммадсолих Саъдин угли
[*]Самаркандский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хоразмий, Самарканд, Узбекистан
muhammadsolih927@gmail.com

Аннотация. В данной статье рассматривается общий подход к описанию и кодированию методов, используемых при включении текстов в национальный корпус узбекского языка. Общий формат может быть оправдан разнообразием и несовместимостью существующих текстовых форматов. Используя формат JSON для хранения текстов в корпусе, можно увеличить скорость поиска в корпусе и преодолеть теоретические и технические проблемы масштабируемости. Описано включение в состав корпуса текстов эпоса «Алпомыш».

Abstract. This article discusses the general approach to the description and coding of the methods used in the inclusion of texts in the national corpus of the Uzbek language. A common format can be justified by the diversity and incompatibility of existing text formats. By using the JSON format to store texts in the corpus, it is possible to increase corpus search speed and overcome theoretical and technical problems of scalability. The inclusion of the texts of the Alpomish epic into the corpus is described.

Ключевые слова: *корпус, форматирование, файл, текст, эпос Алпомыш, токен, разметка, тэг, тэггер, формат JSON, формат DOCX.*

Keywords: *Corpus, formatting, file, text, Alpomish epic, token, markup, tag, tagger, JSON format, DOCX format.*

Последние дни национальные корпуса стали неотъемлемой частью лингвистики, как словари и грамматики. После появления корпуса лингвистические науки изменились, можно сказать, что вся лингвистика стала корпусной лингвистикой. Примеры наиболее известных и признанных лингвистических корпусов: Национальный корпус русского языка (https://ruscorpora.ru/new/), Национальный корпус Великобритании (http://www.natcorp.ox.ac.uk/, https: //www.english-corpora.org/bnc/), Турецкий национальный корпус (https://www.tnc.org.tr/), Американский национальный корпус (http://www.anc.org/) и другие[1].

Тексты доступны в различных форматах PDF, изображений, документов и других форматах. Перед добавлением текстов в корпус необходимо преобразовать имеющиеся текстовые файлы в формат *.docx Microsoft Office версии 2010 и выше. Тексты в других форматах конвертируются в формат *.docx с помощью специальных программ, и при конвертации в формат *.docx исходное состояние текста может быть повреждено. При этом орфографические ошибки в тексте доводятся до уровня исходного состояния текста с помощью ручной работы. После этого текст можно загрузить в корпус. В данном исследовании для хранения текстов в корпусе использовался формат JSON (рис. 1).

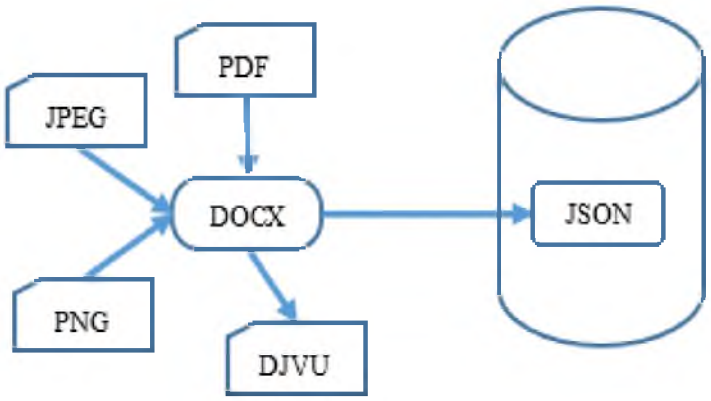


Рисунок 1. *Формат текста и формат корпуса*

Список использованной литературы

1. A.B.Karshiev, S.A.Karimov, M.S.Tursunov, Development of a Modern Corpus of Computational Linguistics // Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT), DOI: 10.1109/ICISCT50599.2020.9351376, 2021.