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.

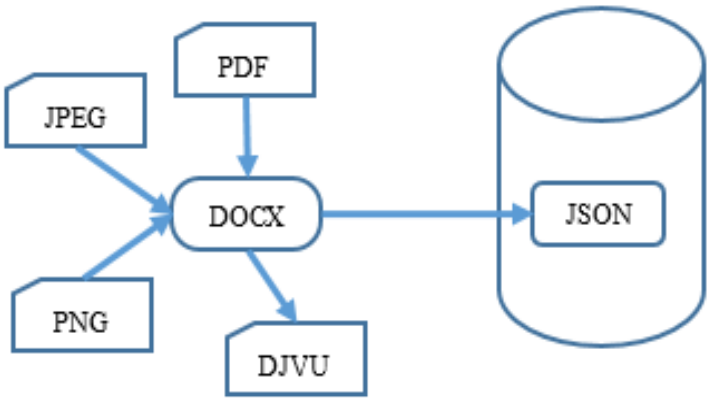
**Abstract**. This article discusses the general approach to the description and coding of the methods used in the inclusion of texts in the national corpus of the Uzbek language. A common format can be justified by the diversity and incompatibility of existing text formats. By using the JSON format 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.