

MUALLIFLIK KORPUSLARINI YARATISH BOSQICHLARI (Usmon Azim dramalarining til korpusi dasturining yaratilishi misolida)

Maxammadiyeva Diyora Raxmatovna

diyoramaxammadiyeva@gmail.com

O'zbekiston-Finlandiya pedagogika instituti xodimi

Annotatsiya: Ushbu maqolada O'zbekiston xalq shoiri Usmon Azimning “Adibning umri” dramatik asarlar to'plamiga kiritilgan dramalari uchun mualliflik korpusining xususiyatlari va undan foydalanishning ahamiyati yoritilgan.

Abstract: This article describes the characteristics of the author's corpus and the importance of its use for the dramas of the national poet of Uzbekistan Usman Azim included in the collection of dramatic works "The Life of an Author"

Аннотация: В данной статье описаны особенности авторского корпуса и значение его использования для драм народного поэта Узбекистана Усмана Азима, включенных в сборник драматических произведений «Жизнь автора».

Kalit so'zlar: *Til korpusi, ta'limiy korpus, dramatik korpus, korpusdan foydalanish, konkordans, korpusning ahamiyati.*

Korpus lingvistikasi kompyuter lingvistikasining yangi tarmog'i sifatida so'nggi yillarda shakllandi va bu sohada bir qancha ilmiy, amaliy loyihalar amalga oshirildi. Har bir korpus, albatta, lingvistik tahrirdan, ishlov berishdan o'tgan va dasturiy ta'minot bilan ta'minlangan bo'lishi lozim. Bugungi kunda korpuslar leksikografiya va grammatika kabi tilshunoslik sohalarining ajralmas qismiga aylandi. Korpus paydo bo'lganidan so'ng tilshunoslik fanlari yangi umr yo'lini boshlashdi, aytish mumkinki, butun tilshunoslikni korpus tilshunosligi materiali sifatida qamrab olish mumkin.

O'rganilgan adabiyotlar natijasidan ma'lum bo'ldiki, korpusni loyihalashtirish jarayonida material (matn) tanlash, saralash, uni texnik jihatdan korpusga moslashtirish eng asosiy bosqichdir [Хамроева, 2018]. Usmon Azim dramalari til korpusi dasturini yaratish ishlarida lingvistik resurs sifatida ijodkorning “Adibning umri” [Usmon Azim, 2015] to'plami tanlandi va u quyidagi bosqichlarda amalga oshirildi:

1. Matnning belgilangan manbaga muvofiq holda korpusga kirishi ta'minlandi.

2. Matn avtomatik o'qilish shakliga keltirildi, ya'ni kiritiladigan matnning elektron shakldagi na'munasi Usmon Azimning “Adibning umri” dramatik asarlar to'plamidan skanerlash yo'li orqali docx formatda hosil qilindi.

3. Tahlil, matnga dastlabki ishlov berish bosqichida skaner qilingan matnlar dastur yordamida kril alifbosi shaklidan lotin alifbosi shakliga o'tkazildi. Shundan so'ng filologik tekshiruv, tahrirdan o'tkazildi. Bunda hosil qilingan matn kitobdagi

namunasi bilan solishtirildi, texnik jarayondagi imloviy, ishoraviy xatolar bartaraf etildi

4. Konversiyalash, grafematik tahlil. Bunda matnlar qayta kodlashtirish jarayoniga qadar mashina ishlovidan o'tkazildi, nomatniy qismlar (rasm, jadval) o'chirildi. Matndagi bo'g'in ko'chirish, chegaralar (MS-DOS matnlarida) bekor qilindi, tire, boshqa belgilar bir xilligiga erishildi. Grafematik tahlil korpusga kiruvchi matnni qismga (so'z, bog'lovchi) ajratish, nomatniy elementni o'chirish kabi amallarni bajarishdan iborat.

5. Mashina ishlovining so'nggi bosqichida nostandart (noleksik) elementni belgilash, rasmiylashtirish, maxsusmatniy elementni (qisqartma asosida yozilgan nom (ism, familiya), boshqa alifboda yozilgan o'zlashma leksema, rasmga berilgan nom, izoh, zarvara,q, adabiyotlar ro'yxati va b.) bir xil mezon asosida qayta ko'rib chiqish ishlari bajarildi. Albatta, bu amallar avtomatik ravishda matn muharriri tomonidan bajariladi.

Endi Usmon Azim dramalari til korpusining dasturiy ta'minot strukturasi haqida to'xtalamiz. Dasturiy ta'minot ikki qismdan iborat bo'ladi (2.1-rasm) [Qarshiyev, Tursunov, 2022]:

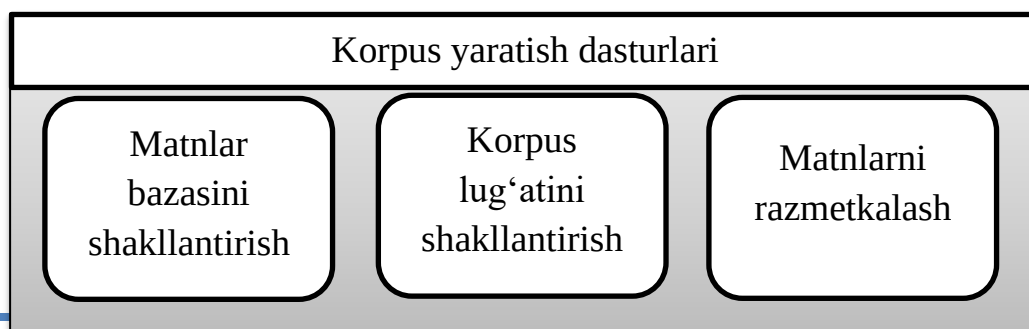
- 1) korpusni yaratishga mo'ljallangan dasturlar;
- 2) korpusdan foydalanishga xizmat qiladigan dasturlar.



2.1-rasm. Dasturiy ta'minot tarkibi

Korpus yaratishga mo'ljallangan dasturlar

Korpusni yaratishga mo'ljallangan dasturlar matnlar bazasini shakllantirish, korpus lug'atini shakllantirish va tahrirlash, matnlarni razmetkalash kabi vazifalarni bajarishi kerak (2.2-rasm) [Qarshiyev, Tursunov, 2022].



2.2-rasm. Korpus yaratish dasturlari tarkibiy qismlari

Matnlar bazasini shakllantirish dasturlari

Shuni eslatib o'tish kerakki, bizning ushbu tadqiqotimiz doirasida yaratilgan mualliflik korpusi ma'lum bir vazifalarni bajaradigan, foydalanuvchilarning ma'lum bir doirasiga yo'naltirilgan maxsus kichik korpus sanaladi.

Usmon Azim dramalari til korpusi dasturining matnlar bazasini shakllantirishda quyidagi ishlar bajarildi:

- matnlarni raqamlashtirish, ularni tahrirlash;
- matn bo'yicha ma'lumotlarni faylga yozish;
- matnni bazaga kiritish.

Matnlarni raqamlashtirishda qog'oz manbadagi matn bo'lganligi uchun dastlab uni skanerlab oldik. Bunda skanerlash apparati va skanerlash dasturlari ishlatildi. Hosil bo'lgan elektron matn *MS Word* dasturi yordamida *.docx formatga o'girildi, so'ngra asosiy manba bilan solishtirilib, sinchiklab tekshirildi, xatolar to'g'rilandi va saqlandi.

Matnlar bazasini yaratishga doir ishning keyingi bosqichlarini *Dast_MtnBaza* nomli dastur amalga oshiradi. Bu dastur saqlangan fayllar asosida korpusning matnlar bazasini shakllantiradi. Matnlar bazasi kompyuter xotirasidagi bitta *Matn_Baza* nomli papka va bitta *MeteRazm* nomli metarazmetka ma'lumotlari faylidan iborat bo'ladi. *Matn_Baza* papkada korpus tarkibiga kiritishga qaror qilingan matnlarni o'z ichiga oluvchi *.docx fayllar saqlanadi. *MeteRazm* faylda *Matn_Baza* papkaga kiritilgan har bir faylda mavjud matn haqida metarazmetka ma'lumotlari saqlanadi. Metarazmetka ma'lumotlari matn bo'yicha umumiy ma'lumotlardan iborat bo'lib, unda matn nomi, matn muallifiga oid ma'lumotlar saqlangan.

Gram_Lugat nomli dastur yordamida korpus lug'ati shakllantirildi. Lug'at korpusning matnlar bazasiga kiritilgan matnlar asosida shakllanadi. *Gram_Lugat* dastur korpusning *Matn_Baza* matnlar bazasidagi har bir faylni navbat bilan ishlovdan o'tkazadi. Har bir fayl uchun *Gram_Lugat* dastur quyidagi ishlarni amalga oshiradi [Qarshiyev, Tursunov, 2022]:

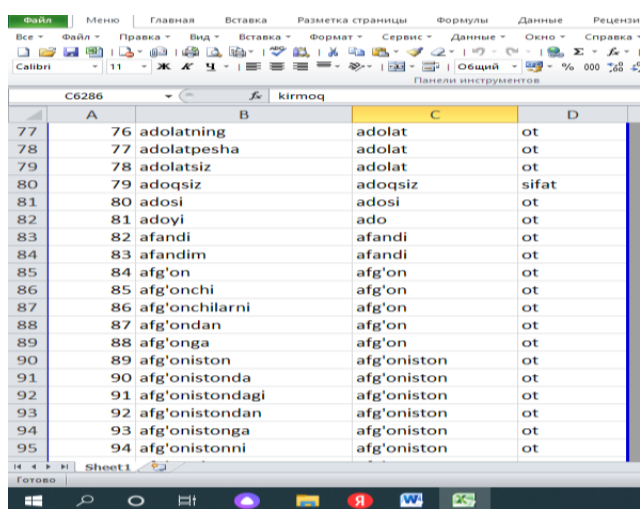
- matnni tokenlash;
- matndagi so'zlarning turli ro'yxatlarini (alfavit-chastotali tartibdagi, chastota-alfavit tartibdagi va ters ro'yxat) tuzish;
- so'zlarni grammatik lug'atga kiritish;

Matnni tokenlashda, matnni avtomatik qayta ishlashda, birinchi navbatda, undagi so'zlarni ajratib olish yoki boshqacha aytganda, matnni birliklarga bo'laklash masalasi yuzaga keladi. Buning uchun ajratuvchi belgilarni (probel, tinish belgilari va h.k.) o'z ichiga olmagan hamma qisman satrlar matndan ajratilib

olinishi lozim. Bu esa tokenlar to'plami bo'ladi. Matnlarni avtomatik qayta ishlashning fundamental algoritmlaridan biri, berilgan matnni tokenlarga bo'lib tashlashdan iborat. Algoritm kirishiga matn berilib, chiqishida matndagi tokenlar ro'yxati olinadi. Bu algoritmnini amalga oshiruvchi dasturni tokenayzer deb atashadi. Odatda, tokenlar so'z shakllari bilan bir xil ma'noni beradi. Lekin, leksik birliklarni ifodalash uchun “so'z” emas, balki “token” termini ishlatiladi. Bunga sabab, ba'zi hollarda token sifatida so'zdan kichikroq birliklar (alohida morfema) yoki so'zdan kattaroq birliklar (so'z birikmalari) ishlatilishi mumkin.

Tokenayzer matnni, dastlab, so'zlar orasidagi probellar (bo'shliq belgilari) asosida bo'laklaydi, so'ngra so'zlardan tinish belgilari olib tashlanadi. Qisqartmalar (masalan, TATU, BMT, MDH, h.k.) va sana yozuvi (masalan, 09.04.2018) ham token sifatida olinadi.

Tokenayzer dramalar matnini tokenlarga bo'laklab berdi va dastur yordamida so'zshakllardan alfavitli-chastotali lug'at hosil qilinib, quyidagi rasmda keltirilganidek morfologik razmetka ishlari amalga oshirildi (2.3-rasm).



	A	B	C	D
77	76	adolatning	adolat	ot
78	77	adolatpesha	adolat	ot
79	78	adolatsiz	adolat	ot
80	79	adoqsiz	adoqsiz	sifat
81	80	adosi	adosi	ot
82	81	adoyi	ado	ot
83	82	afandi	afandi	ot
84	83	afandim	afandi	ot
85	84	afg'on	afg'on	ot
86	85	afg'onchi	afg'on	ot
87	86	afg'onchilarni	afg'on	ot
88	87	afg'ondan	afg'on	ot
89	88	afg'onga	afg'on	ot
90	89	afg'oniston	afg'oniston	ot
91	90	afg'onistonda	afg'oniston	ot
92	91	afg'onistondagi	afg'oniston	ot
93	92	afg'onistondan	afg'oniston	ot
94	93	afg'onistonga	afg'oniston	ot
95	94	afg'onistonni	afg'oniston	ot

2.3-rasm. Morfologik annotatsiyalash.

Bu dasturlar quyidagi sifatlarga ega bo'lishi kerak:

- foydalanuvchi so'zlar yoki gaplarning grammatik yoki sintaksis xarakteristikalarini yetarli darajada to'liq aniqlash va o'zgartirish imkoniyatiga ega bo'lsin;
- tahlil maksimal darajada qulay va sodda bo'lib, tez bajarilsin va bu tahlilning to'liqligini kamaytirmasin.

Korpus yaratishda quyidagi savollarga asoslangan holda asosiy talablarni belgilab olish muhim [Qarshiyev, Tursunov, 2022]:

1. Korpus foydalanuvchisi kim? (Individual, guruh, lingvistik jamiyat).
2. Korpusning mantiqiy g'oyasi nimada?
3. Korpusni qurishda qancha hajmdagi ma'lumotlar bilan ishlamoqchisiz?



4. Nima ishlatiladi: matn parchalari, to'liq matnlar yoki ikkalasi ham?
5. Korpusdagi matnlarni tanlash tartibi qanday?

Usmon Azim dramalari til korpusi lingvistik korpuslar qo'llanilishiga qo'yilgan quyidagi talablar bajaradi:

1. Korpus o'quv va ma'lumot qo'llanmalar yozishda, lug'atlar tuzishda hamda lingvistik ekspertizalar va tahlillarni ishlab chiqishda, til ma'lumotlari sifatida ishonchli manba hisoblanadi;
2. Korpus turli xil lingvistik fanlarni o'qitishda xuddi tilning ko'rgazmali (illyustratsiya) manbasi, hamda tahlillarni taqdimotlilik ko'rinishi sifatida ishlatish mumkin;
3. Korpus ilmiy tadqiqot ishlarida.
4. Korpus til faoliyatining ma'lumotlar qo'llanmasi sifatida bo'lishi mumkin.

Xulosa qilib aytganda, razmetkaning yuqori aniqlikda bo'lishini faqat filolog mutaxassis ta'minlashi mumkin. Filolog mutaxassis har bir matnning razmetkalanishini so'zma-so'z sinchiklab tekshirishi, avtomatik razmetka xatolarini tuzatishi va omonim holatlarni bartaraf qilishi kerak. Ishning asosiy qismi qo'l mehnati orqali bajarilishi mumkin. Shuning uchun filolog mutaxassisning kompyuterdagi ishiga mo'ljallangan maxsus razmetkalash dasturlari yaratilishi zarur.

Foydalanilgan adabiyotlar:

1. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. filol.fan.bo'y. PhD diss. –Qarshi, 2018.
2. Usmon Azim. Dramatik asarlar. - Toshkent, 2015.
3. A.Qarshiyev, M.Tursunov. O'zbek tili milliy korpusini loyihalash. "O'zbek milliy va ta'limiy korpuslarini yaratishning nazariy hamda amaliy masalalari", 2022.
4. A.Qarshiyev, M.Tursunov. O'zbek tili milliy korpusining dasturiy ta'minot strukturasi va vazifalari, 2022.