Academic Research in Educational Sciences
ISSN: 2181-1385
DOI: 10.24412/2181-1385-2022-3-644-650

Volume 3 | Issue 3 | 2022
Cite-Factor: 0,89 | SIS: 1,12
SJIF: 5,7 | UIF: 6,1

O‘ZBEK TILI ELEKTRON KORPUSIDA (http://uzbekcorpus.uz/) OG‘ZAKI MATNLAR KORPUSINI YARATISHNING NAZARIY VA AMALIY MASALALARI

Nilufar Zaynobiddin qizi Abduraxmonova
Filologiya fanlari doktori (DSc), O‘zMU dotsenti
abdurahmonova.1987@mail.ru

Mavluda Yangiboyevna Urazaliyeva
O‘zMU kompyuter lingvistikasi I bosqich magistranti
mavludaurazalieva@gmail.com

ANNOTATSIYA

Maqolada zamonaviy kompyuter texnologiyalar yordamida yaratilgan korpus va uning imkoniyatlarini takomillashtirishga oid fikrlar tahlilga tortilgan. O‘zbek tili elektron korpusining audiomatnli korpuslarini yaratishda xorij tajribasi o‘rganilib, og‘zaki nutq aktlarini korpus bazasiga kiritishning amaliy jihatlariga e’tibor qaratilgan.

Kalit so‘zlar: elektron korpus, kompyuter lingvistikasi, audiomatnli korpus, multimediya, o‘zbek tili

ABSTRACT

The article provides information about the corpus, which was created using modern computer technologies, and carried out a wide range of works to improve its capabilities. It focuses practical sides input database of speech act into Uzbek electronic corpus.

Keywords: electronic corpus, computational linguistics, audio corpus, multimedia, Uzbek

KIRISH

I. Korpusshunoslikning o‘rganilishi

Tilshunoslik va boshqa sohalarda elektron korpusdan foydalanish XX asrning ikkinchi yarmiga kelib aynan kompyuter lingvistikasi rivojlanish bosqichida yanada keng tus oldi.

Tilning lisoniy muammolarini o‘rganish va hal qilishda kompyuter texnologiyalardan foydalanish ma’lumotlarning

elektron bazasini yaratishda katta yordam beradi. Buning uchun korpus yaratish va unga tilda mavjud so‘zlar, iboralarni kiritib elektron bazasini yaratish muhim.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Dunyo tajribasida korpus yaratishning lingvistik, matematik va dasturiy jihatlari olimlar tomonidan qilingan bir qancha ishlarda o‘z ifodasini topgan.[1] Chunonchi, rus va ingliz tillari bo‘yicha korpus lingvistikasi turli sohalar kesimida V.Zaxarov, A.Sedov, A.Baranov, R.Potapova, V.Rikov, U.Frensis, N.Leontyeva, V.Martin, S.Kubler, A.Laurans, E.Etwell, S.Hunston, L.Boizou, McKenneri, J.Grafmiller, J.Grieve, N.Grum, S.Hansson, K.McAulif, M.Malberg, P.Milin, A.Murakami, R.Peych, A.Shembri, P.Tompson, B.Vinter, G.Lich kabi xorijiy olimlar tomonidan ham turkologiyada korpus lingvistikasi bo‘yicha ilmiy tadqiqotlar olib borilgan.

O‘zbek korpus lingvistikasi bo‘yicha Sh.Hamrayeva, N.Abduraxmonova, Sh.Gulyamova, G.Toirova kabi olimlarning ishlarini keltirish o‘rinlidir.

Korpus ko‘plab yo‘nalishlarda kompyuter lingvistikasi, tarjimashunoslik, pedagogika kabi sohalarning tadqiqot obyekti vazifasini bajargani bois mazkur sohada olib borilayotgan ishlarning oxirgi o‘n yillikda sezilarli darajada oshdi.

II. Korpus tadqiqida yondashuvlar tahlili

Manbalarga ko‘ra 1990-yilga kelib dunyo tillarining kompyuter tahliliga mo‘ljallangan 600 ga yaqin korpusi borligi aniqlangan[2].

Istalgan tildagi audiokorpusni yaratishda, avvalo, barcha uslublardagi katta hajmga ega bo‘lgan elektron manba, ularning audiomatni bo‘lishi kerak. Uning interfeysida *izlash* buyrug‘i yosh, jins, millat, davr va boshqa jihatlar bo‘yicha qidirish imkoniyati mavjud. Bunday korpuslar tilshunoslikning turli sohalarida xususan, *lingvodidaktika, qiyosiy tilshunoslik, tarjima* sohalarda katta yordam beradi. Zero, xususiy auditoriyaga tegishli audiomatn foydalanuvchilar uchun juda qulay va tilni o‘rganing samarali usuli hamdir.

Dunyoda Multimediyali rus tili korpusi (МУРКО), Yevropa Ittifoqi korpusi asosida ko‘ptilli korpus (ECI/MCI), Ingliz milliy korpusi (BNC)larda mavjud audiokorpuslar yaratilgan. Ular orasida mashhur yozuvchi va shoirlarning mualliflik

[1]Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari (monografiya) Toshkent, 2021. – B. 7-8.
[2]Захаров В.П., Богданова С.Ю Корпусная лингвистика: учебник для студентов гуманитарных вузов, Иркутск, ИГЛУ, 2011 – С.12.

korpuslar ham mavjud. A.P.Chexov, U.Shekspir, Dante, A.S.Pushkin kabilarning ijodiga bag‘ishlangan mualliflik korpuslaridan audiokorpuslar ham o‘rin egallagan.

Ilk bor Factored va MLCommons tomonidan MSWC – Ko‘p tilli og‘zaki so‘zlar korpusining birinchi versiyasi yaratildi. Bu korpus 50 xil tildagi katta hajmdagi ovozli ma’lumotlarni o‘z ichiga oladi. Bu tillarda 5 milliarddan ortiq kishilar so‘zlashadi va ko‘pgina tillar uchun bu ovozli interfeys ta’lim olish uchun mo‘ljallangan ilk cheklanmagan bepul ma‘lumotlar bazasidir.

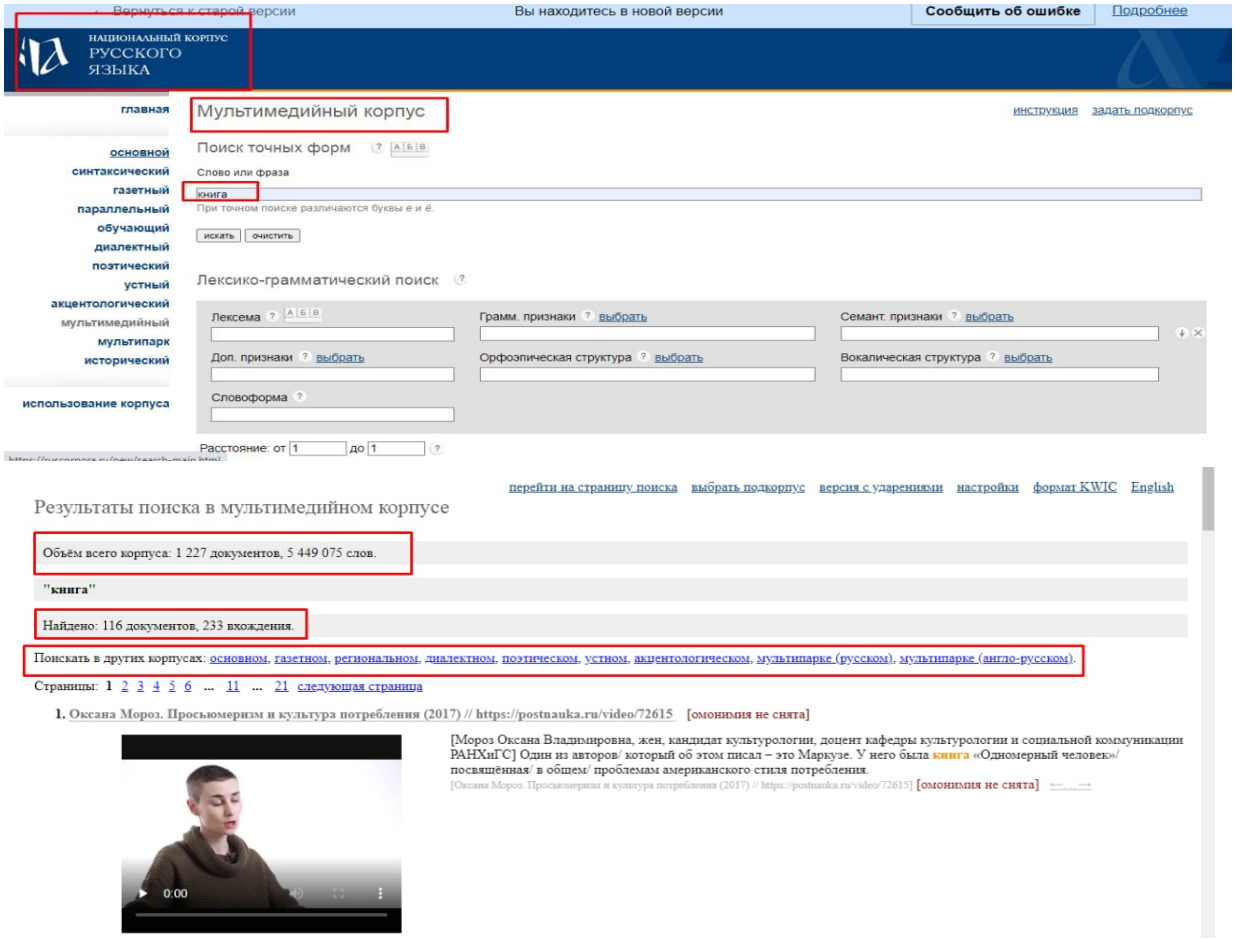
Kalit so‘zlarni aniqlash, og‘zaki termin orqali qidirish va turli sohadagi odamlarga foyda keltiruvchi boshqa dasturlar sohasidagi akademik tadqiqotlarni va tijorat ishlarda foydalanishga mo‘ljallangan. Bunda har qanday tildagi kalit so‘zlar uchun ovozli interfeys yaratish maqsad qilib qo‘yilgan.

Ovozli dasturlar allaqachon kundalik hayotga kirib kelgan. Masalan, foydalanuvchi atrofidagi holatlarni aniqlash ko‘plab aqlli ilovalar (masalan, Apple Siri, Amazon Alexa yoki Google ovozli yordamchisi) zimmasiga yuklatilgan. Chiroqni o‘chirish yoki murakkabroq interfeysni ishga tushirish kabi harakatlarni boshqarishda buyruq ohangidagi so‘zlarni to‘xtovsiz eshitish uchun kalit so‘zlarni aniqlash tizimi yaratilgan. Bunday ovozli dasturlar ba’zi odamlar uchun axborot asrida qulaylik hisoblansa, ko‘zi ojiz kishilar uchun muhim ta’lim olish vositasi hamdir.

Bunday dasturlar katta ma’lumotlar bazasining kompyuter modellarini o‘rganishni talab qiladi. Aslida korpus bunday dasturiy ta’minot uchun kalit so‘zlar turli kontekstlardagi minglab so‘zlarni to‘plash va tekshirish uchun resurs bo‘lib xizmat qiladi. MLCommons MSWC 50 ta tildagi nutqni aniqlash uchun katta hajmdagi ma’lumotlar bazasini yaratishda tabiiy tilning audiomatnli korpusidan foydalanmoqda va u doimiy ravishda yangilanib boradi. Umuman olganda, ma’lumotlar bazasi 340 000 dan ortiq so‘zni va 6000 soatdan iborat 23 million miqdordagi bir daqiqali audiomatnlarni o‘z ichiga oladi. Ushbu ma’lumotlar to‘plamining ochiq manbali resurslarini yaratishda foydalanuvchilar takliflarida mavjud alohida so‘zlarini ham ajratib uchun qo‘llaniladi. Bu esa turli tillarda ovozli yordamchilar uchun kalit so‘zlarni aniqlash modellarini o‘qitish uchun ishlatilishi mumkin.

MSWC da ma’lumotlar bazasidagi tillardan 12 tasi eng ko‘p qo‘llaniladigan 100 soatdan ortiq ma’lumotlar, 12 tasi 10 soatdan 100 soatgacha bo‘lgani o‘rtacha ishlatiladigan ma’lumotlar va 26 tasi kam ma'lumotli 10 soatdan kam bo‘lgan kam qo‘llaniladigan tillardir. MSWC ma’lumotlar to‘plami ushbu tillardan 46 tasi uchun ochiq manbali og‘zaki nutq

ma’lumotlarining yagona to‘plamidir. Har bir kalit so‘zni o‘rganish, tekshirish va test qilish uchun oldindan belgilangan bo‘linmalarga ega va ma’lumotlar bazasini yaratish va kalit so‘zlarni tasniflash uchun ishlatiladigan ochiq manba vositalarini ham chiqarish mumkin.



Rus milliy korpusidagi multimediyali korpusning hajmi 5 449 075ta so‘zni tashkil qiladi. Mazkur subkorpus doimiy ravishda yangilanib boradi. Har bir berilgan video 8-30 soniyalarda aks etgan. Har bir tovush ohangi, unlilar talaffuzi alohida-alohida keltiriladi. Har bir uslubdan olingan matn va audiolardagi ovoz egasining yoshi, jinsi, millati ko‘rsatiladi. Bu esa dialektologiya uchun juda zarur va juda muhim manba bo‘lib xizmat qiladi.

III. O‘zbek tilining audiomatnlar korpusini yaratishning amaliy masalalari

O‘zbek tilida yuqoridagi kabi audio dasturlar va korpuslar yaratishda tilimizda mavjud beshta uslub (so‘zlashuv, badiiy, ilmiy, rasmiy, publitsistik)larga oid manbalarni yig‘ib, elektron matni va ovozli variantini davr, yosh, jins, soha vakillariga qarab guruhlab chiqilsa o‘zbek milliy korpus rivojiga katta hissa qo‘shiladi.

Shunisi ahamiyatliki, ovozli matn korpusining lingvistik bazasi muayyan yozma matnni o‘qish natijasida yaratiladimi yoki televideniye, jonli muloqot, radiodagi diolog yoki monologlardan tuziladimi degan masala birlamchi sanaladi. Shuningdek, audiomatnni transkripsiya qilish va standart tilda raqamlashtirish eng muhim bosqich sanaladi. Og‘zaki nutqning etnografik ma’lumotlarni metama’lumot sifatida berish ham muayyan darajada asosiy talablardan biri. So‘zlovchining qaysi hududga tegishli ekanligi, jinsi, yoshi, dialekti, kasbi bularning hammasi audiomatnni korpusga kiritishda bosh mezon sanaladi.

N.Abdurahmonovaning “O‘zbekcha matnlarni ovozlashtirish dasturining lingvistik ta’minotini ishlab chiqishda ayrim masalalar tadqiqi” nomli maqolasida so‘z turkumlari, tinish belgilari, arab va rim raqamlarini yozish va o‘qishda uchrovchi bir qator kamchiliklar sifatida keltiriladi. Bunda bazaga ma’lumot kiritishda matnning qaysi bandida *chiziqcha*, qaysi birida *tire* ekanligi va *-inchi* qo‘shimchalariga ham e’tiborli bo‘lish kerak. Yaratiladigan dastur esa buni tushunib olishi lozim. Tinish belgilari yozilgan paytda qo‘yiladigan belgilar ovozli matnda o‘qilmaydi. O‘zbekcha matnlarni ovozlashtirish dasturining har qanday o‘zbek tilidagi matnlarni hech qiyinchiliksiz o‘qib berishda uning lingvistik ta’minotining qay darajada muakammal ishlab chiqilgani katta ahamiyatga egadir. Shuningdek, o‘zbek tiliga boshqa tillardan, asosan, rus tili va u orqali boshqa tillardan o‘zlashgan ruscha internatsional so‘zlarni tadqiq etish va bunday so‘zlarni dastur lingvistik ta’minotiga kiritish masalalarini o‘rganish vazifasi ham oldimizda ko‘ndalang turibdi. O‘zlashma so‘zlarning talaffuzi o‘zbek tili so‘zlari talaffuzidan farq qilgani bois ham ularning audio formatdagi va yozma shaklini lingvistik ta’minotga kiritish dasturning bunday so‘zlarni xatosiz o‘qishiga imkon yaratadi[3].

XULOSA

Bundan tashqari o‘zbek tilida unli harflarning qisqa cho‘ziqligi ham og‘zaki nutqda ta’sir etmay qolmaydi: *ilm* [il:m]⇔*bilim [*bılım] *o‘lim* [olka] ⇔ *o‘lka* [ölka] kabilar.

Iste’moldan chiqish xavfi ostiga kelib qolgan tillar uchun ularning elektron bazasi va korpusini yaratish, shu tilga taalluqli bo‘lgan ilmiy va badiiy adabiyotlar yillar davomida asrashga, ular ustida bir qancha ilmiy ishlar qilishga imkon beradi.

Umuman olganda, audio korpuslar ta’limga ayniqsa, maktab yoshidagi bolalar nutqini kuzatib borishda yuqori samaradorlikka erishishga yordam

[3] N.Abduraxmanova O‘zbekcha matnlarni ovozlashtirish dasturining lingvistik ta’minotini ishlab chiqishda ayrim masalalar tadqiqi // Turklang-2018 xalqaro konferensiya, Toshkent, 2018.

beradi. Sababi til ijtimoiy hodisa sifatida doimiy ravishda o‘zgarib turadi, qaysidir so‘zlar neologizm sifatida kirib kelsa, ba’zilari esa tarixiy so‘zlarga aylanadi. Bu jarayonni esa multimediali korpus orqali bevosita kuzatib borish mumkin. Ko‘rinib turibdiki, korpus nafaqat soha kishilarning, balki tilni rivojlantirishda umummilliy masala hisoblanadi.

REFERENCES

1.Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In: Abstract of the VI International Conference “Modern Problems of Applied Mathematics and Information Technology - Al-Khorezmiy 2018”, pp. 37–38, Tashkent, Uzbekistan (2018)

2.Abduraxmonova, N. Z. "Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref." (2018).

3.Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016;2 (38):12-7.

4.Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education. 2018(3):68.

5.Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.

6.Abdurakhmonov N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research. 2017;5(03):89-100.

7.Kubedinova L. Khusainov A., Suleymanov D., Gilmullin R., Abdurakhmonova N. First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020) .2020/11: 90-101

8.Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.

9. N. Abdurakhmonova, U. Tuliyev and A. Gatiatullin, "Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz," 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670043.
10. https://ruscorpora.ru/new/
11. https://www.english-corpora.org/bnc/
12. http://www.turklang.net
13. https://mlcommons.org/en/multilingual-spoken-words/