KOMPYUTER LINGVISTIKASIDA AUDIOKORPUS MASALALARI

Urazaliyeva Mavluda
Mustaqil izlanuvchi (PhD)
O‘zbekiston Milliy universiteti.

Annotatsiya: Maqolada zamonaviy kompyuter texnologiyalar yordamida yaratilgan korpus va uning imkoniyatlarini takomillashtirishga doir bir qancha 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: kompyuter lingvistikasi, elektron korpus, audiokorpus, multimediyali korpus, o‘zbek tili.

Аннотация: В статье анализируется корпус, созданный с помощью современных компьютерных технологий, и несколько идей по улучшению его возможностей. Изучен зарубежный опыт создания аудиотекстовых корпусов электронного корпуса узбекского языка, а также уделено внимание практическим аспектам включения устных речевых актов в базу корпуса.

Ключевые слова: компьютерная лингвистика, электронный корпус, аудиокорпус, мультимедийный корпус, узбекский язык.

Abstract: The article analyzes a corpus created using modern computer technologies and several ideas for improving its capabilities. Foreign experience in creating audio-text corpora of the electronic corpus of the Uzbek language has been studied, and attention has also been paid to the practical aspects of including oral speech acts in the corpus database.

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

Tilshunoslik va boshqa sohalarda elektron korpusdan foydalanish kompyuter lingvistikasi rivojlanish bosqichi davriga kelib, aynan XX asrning ikkinchi yarmida yanada keng tus oldi.

Tilning lisoniy muammolarini o‘rganish va ularga yechim topishda kompyuter texnologiyalardan foydalanish ma’lumotlarning elektron bazasini yaratishda katta yordam beradi. Buning uchun korpus yaratish va unga tilda mavjud so‘zlar elektron bazasini yaratish muhim.

Dunyo tilshunosligida korpus yaratishning lingvistik, matematik va dasturiy jihatlari olimlar tomonidan qilingan bir qancha ishlarda o‘z ifodasini topgan [Abduraxmonova N., 2021; 7-8.]. 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.

Ilk bor Factored va MLCommons tomonidan MSWC – Ko‘p tilli og‘zaki so‘zlar korpusining birinchi versiyasi yaratildi [https://mlcommons.org/en/multilingual-spoken-words/]. 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. MSWCda kalit so‘zlarni aniqlash, og‘zaki termin orqali qidirish va turli sohadagi odamlarga foyda keltiruvchi boshqa dasturlardagi akademik darajadagi tadqiqotlarni va tijorat ishlarida foydalanishga mo‘ljallangan. Bunda har qanday tildagi kalit so‘zlar uchun ovozli interfeys yaratish maqsad qilib qo‘yilgan.

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.

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. Hozirda mavjud ovozli platformalardan biri bo‘lgan 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 [https://mlcommons.org/en/multilingual-spoken-words/].

Rus milliy korpusidagi multimediyali korpusning hajmi 5 763 881 ta so‘zni tashkil qiladi [https://ruscorpora.ru/]. 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.

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 [Abduraxmonova N., Turklang, 2018].

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* [olim] ⇔ *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.

Ma’lumotlar bazasida berilgan so‘zlar orfografik jihatdan to‘g‘ri yozilishi ham hisobga olinishi kerak. Natoijada morfoanalizator to‘g‘ri va sifatli ishlashiga zamin yaratilgan.

Korpusning kundalik hayotdagi ahamiyati yetarlicha bo‘lib, uni avtomatik tarjimada sintaktik tahlil uchun yordam, audio korpusda matnn tahlil qilish uchun baza, qolaversa, multimediyali darslarda elektron doskalarda til o‘rgatishda keng qo‘llash mumkin.

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

Foydalanilgan adabiyotlar:

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. 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.
3. Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.
4. 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, https://doi.org/10.1109/ICISCT52966.2021.9670043
5. https://ruscorpora.ru/new/
6. http://www.turklang.net
7. https://mlcommons.org/en/multilingual-spoken-words/