TURKIY TILLAR KORPUSINING QIYOSIY TADQIQI

Durdona Gurbanmurat qizi Allaberdiyeva
Mirzo Ulug‘bek nomidagi O‘zbekiston milliy universiteti magistranti
durdonallaberdiyeva39@gmail.com

ANNOTATSIYA

XXI asrning eng dolzarb ijtimoiy masalalaridan biri tabiiy tillarni saqlab qolishdir. Dunyo tillarining elektron korpuslarini yaratish va rivojlantirishda NLP va til texnologiyalariga doir tadqiqotlarni izchil ravishda olib borish dolzarb masalaga aylandi. Ushbu maqolada turkiy tillarda yaratilgan korpuslar bir-biriga qiyoslash orqali o‘rganiladi va ular haqida ma’lumot beriladi.

Kalit so‘zlar: NLP, korpusshunoslik, dasturiy ta’minot, model, tokenayzer, matnlar bazasi, lug‘at, grammatik ma’lumot.

Dunyo tajribasida korpus yaratishning lingvistik, matematik va dasturiy tomonlari olimlar tomonidan tadqiqotlarda o‘z ifodasini topgan. Chunonchi, rus va ingliz tillari bo‘yicha korpus lingvistikasi sohalar kesimida V.Zaxarov, A.Sedov, A.Baranov, R.Potapova, V.Rikov, U.Frensis, N.Leontyeva, V.Martin, S.Kubler, A.Laurens, E.Etwell, S.Hunston, L.Boizou, Me.Kenneri, J.Grafmiller, J.Grieva, N.Grum, S.Hansson, K.Me.Aulif, M.Malberg, P.Milin, A.Murakami, R.Peych, A.Shembri, P.Tompson, B.Vinter, G.Lich, kabi xorijiy olimlar tomonidan hamda Turkologiyada Korpusshunoslik bo‘yicha (korpus lingvistikasi) sohasi bo‘yicha ilmiy tadqiqotlar olib borilgan. Turk tili koprpusi bo‘yicha Aksan, Deniz, Zeyrek, Kemal, Oflazer, Umut O‘zge Bular; uyg‘ur tili bo‘yicha Yusup Aibaidulla, Kim-Teng Lua; boshqird tili bo‘yicha J.Suleymanov, A.Gatiatullin, O.Neyzorova, R.Gilmullin, B.Hakimov; qirimtatar tili bo‘yicha L.Kubedinova hamda tuva tili bo‘yicha Salchak kabi olimlarning ishlari diqqatga sazovor[1].

Bugungi kunda Turkologiyada korpusshunoslik sohasi turkiy tillar korpuslari bo‘yicha olib borilayotgan tadqiqotlar sababli rivojlanmoqda. Turkiy tillar orasida ham korpusshunoslik bo‘yicha bir qator tadqiqotlar amalga oshirilgan. Xususan, N.Yoqubova, M.Ayimbetov, S.Rizayev, S.Muhamedov kabi olimlarimizning tadqiqotlarini alohida ta’kidlash joizdir. Shuningdek, so‘nggi o‘n yillikda kompyuter lingvistikasi sohasida A.Po‘latov, S.Muhammedova, A.Rahimov, Z.Xolmanova,

[1] Nilufar Abduraxmmanova. O’zbek tili elektron korpusining kompyuter modellari. Monografiya.GlobeEdit,2021.

N.Abdurahmonova kabi olimlarning nazariy va amaliy tadqiqotlari yaqqol ko‘zga tashlanadi.

Qozoq tili korpusi A.Baytursun o‘g‘li nomidagi Tilshunoslik instituti Amaliy tilshunoslik bo‘limi tomonidan yaratilgan bo‘lib, asoschisi A.Jubanov hisoblanadi. Korpus veb-saytida qozoq tilining elektron matn fondi mavjud. Korpusdagi matn hajmi 31 mln. Matnlar qozoq tilining 5 ta uslubidan (badiiy uslub, ilmiy uslub, publitsistik uslub, rasmiy uslub, so‘zlashuv uslubi) to‘plangan. Korpus so‘z, so‘z shakli (so‘zni o‘zgartirish) bo‘yicha qidirish va qidirilayotgan so‘z ishlatiladigan jumlalar ro‘yxatini va ularning manbasini ko‘rish mumkin. Har qanday topilgan so‘z, so‘z shakli yoki misollardagi har qanday so‘z tilning barcha darajalariga tegishli ma'lumotlar bilan ta'minlangan.

Turk tilining milliy korpusi (TUD) 50 million so‘zdan iborat umumiy maqsadli ma’lumotnoma bo‘lib, turli soha va janrlardan bugungi turk tilining yozma va og‘zaki misollarini o‘z ichiga oladi hamda keng qamrovli, platformadir. Foydalanuvchilar o‘z so‘rovlarini cheklash mezonlarining keng doirasi (ommaviy axborot vositalari, matn namunasi, mavzu maydoni, lotin matn formati, muallif jinsi, maqsadli o‘quvchi va matn turi va boshqalar) bilan bajarishlari mumkin. Bundan tashqari, TUD 3.0 versiyasida foydalanuvchilar so‘z turi va qo‘shimchalari bo‘yicha qidirish imkoniyatiga ega.

Tatar tili yozma korpusi “Tugen Tel” Amaliy semiotika instituti tomonidan (2012-2024) yaratilgan bo‘lib asoschisi Sayxunov M.R. (filologiya fanlari nomzodi, Tatariston FA Informatika instituti ilmiy xodimi). Korpusning hajmi 180 000 000 tokenni tashkil etadi (2018 yil dekabrgacha). Korpus turli uslub va janrdagi matnlarni (badiiy adabiyot, ommaviy axborot vositalari matnlari, rasmiy hujjatlar, o‘quv va ilmiy adabiyotlar va boshqalar) o‘z ichiga oladi. Korpusda barcha mavjud grammatik so‘z shakllarini taqdim etishga qaratilgan grammatik izohlar tizimi mavjud. Tatarcha so‘zning grammatik annotatsiyasi so‘zning nutq qismi va morfologik xususiyatlar (parametrlar) to‘plami haqidagi ma'lumotlarni o‘z ichiga oladi. Korpus matnlarini morfologik izohlash PC-KIMMO dasturiy vositasida amalga oshirilgan tatar tilining ikki darajali morfologik tahlil moduli yordamida amalga oshiriladi. Korpusning qidiruv tizimi leksemalar, so‘z shakllari va individual grammatik parametrlarni qidirish imkonini beradi.

Korpus lingvistikasi korpus yaratish hamda korpus metodlaridan foydalanib, tilning nazariy va amaliy muammolarini o‘rganishga doir ikki yo‘nalishi asosida ish olib boradi. Mazkur sohada Respublikamizning bir qator oliy ta’lim muassasalari, shuningdek, ilmiy tadqiqot institutlarida ilmiy izlanishlar olib borilmoqda. O‘zbek korpus lingvistikasi

rivojlanishiga hissa qo‘shayotgan olimlar sirasiga B.Mengliyev, N.Abdurahmonova, Sh.Shahobiddinova, Z.Xolmanova, S.Karimov, L.Raupova, Sh.Hamroyeva, G.Toirova, J.Djumabayeva, G.Ergasheva, A.Eshmo‘minovlarning tadqiqotlarini alohida qayd etish o‘rinli[2].

O’zbek tili korpusi (uzbek.corpora.uz) O’zbekiston milliy universiteti “Kompyuter lingvistikasi va amaliy tilshunoslik” kafedrasi professori N.Abduraxmonova tomonidan yaratilgan, bugungi kunda 30 million so’zdan iborat, lug’at va foydali saytlarni ichiga qamrab olgan keng qamrovli platformadir.

Yuqorida sanab o‘tilgan turkiy tillar korpusi turkiy tillarning ba’zi umumiy grammatik xususiyatlari sababli model jihatdan bir birlariga o‘xshash tarzda yaratilgan.

Xullas zamonaviy leksikografiya va lingvistik tadqiqotlar tarixi, umuman olganda, leksik hujjatlar ma'lumotlar bazalarining evolyutsiyasidan ajralmasdir, chunki biz uchun til tarixi va uning lug'ati haqida yagona ma'lumot manbai matnlardir. Shu sababli bugungi kunda tabiiy tillarimizni saqlab qolishning yagona usuli bu ularni elektronlashtirishdir. Endilikda korpus lingvistikasi tilshunoslikning ajralmas qismiga aylandi. Negaki kompyuter lingvistikasi, sotsiolingvistika, pedagogika, tarjimashunoslik, diskurs analiz kabi sohalarda korpuslardan unumli foydalanilib, ijobiy natijalarga erishib kelinmoqda.

[2] Abduraxmanova N. O‘zbek tili elektron korpusining kompyuter modellari. (monografiya). GlobeEdit, 2021.

REFERENCES

1.Andrew Wilson, Dawn Archer, Paul Rayson Language and computers studies in practical linguistics No 56 / Corpus linguistics around the world New York, 2006. 242 p.

2.Anke Ludeling, Merja Kyoto Corpus linguistics: An international handbook, Volume 2 Berlin Volter de Gruyter, 2009. 606 p.

3.Anke Lüdeling, Merja Kytö Corpus Linguistics An International Handbook,Vol. 1, Berlin, New York: Walter de Gruyter. 2008. 81(2), —P. 246–247 DOI: 10.1080/00393270903392342

4.N. Abdurakhmonova, I. Alisher and R. Sayfulleyeva, "MorphUz: Morphological Analyzer for the Uzbek Language," 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 61-66, doi: 10.1109/UBMK55850.2022.9919579.

5.N. Abdurakhmonova, I. Alisher and G. Toirova, "Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing," 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 73-75, doi: 10.1109/UBMK55850.2022.9919521.

6.N. Z. Abdurakhmonova, A. S. Ismailov and D. Mengliev, "Developing NLP Tool for Linguistic Analysis of Turkic Languages," 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Yekaterinburg, Russian Federation, 2022, pp. 1790-1793, doi: 10.1109/SIBIRCON56155.2022.10017049.

7.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), Tashkent, Uzbekistan, 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670043.

8.D. B. Mengliev, N. Abdurakhmonova, D. Hayitbayeva and V. B. Barakhnin, "Automating the Transition from Dialectal to Literary Forms in Uzbek Language Texts: An Algorithmic Perspective," 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, 2023, pp. 1440-1443, doi: 10.1109/APEIE59731.2023.10347617.

9.Juravskiy D., James H. Martin Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2007 –P. 12-13-P.140.

10.Abduraxmonova N. O‘zbek tili elektron korpusining lingvistik va dasturiy ta’minoti.Maqola.Toshkent-2021

11.Qarshiyev A. B, Karimov S.A, Tursunov M.S. O‘zbek tili milliy korpusining dasturiy ta’minot strukturasi va vazifalari.Maqola.Toshkent-2022.
12.Altay Guvenir and Kemal Oflazer Using a corpus for teaching Turkish morphology Proceedings of the seventh twente workshop on language technology. The Netherlands 16-17 June, 1994. – P. 28-40.
13.Antoinette Renouf, Andrew Kehoe Corpus linguistics: Refinements and reassessments New York, 2009. 471 p.
14.Bern Heine, Heiko Narrog The Oxford handbook of linguistic analysis. / Dougles Biber Corpus-based and Corpus-driven analysis of language variation and use UK: Oxford university, 2015. –P. 193.
15.Hines, T. C., Harris, J. L. and Levy, C. L. An Experimental Concordance Program, Computers and the Humanities 4(3): 161–71.
16.McEnery, T., Xiao, R. and Tono, Y. Corpus-based Language Studies: An Advanced Resource Book. London: Routledg, 2006.
17.Dildor Otajonova.Korpus lingvistikasi tarixiga nazar // Ta’lim jarayonida raqamli texnologiyalarni joriy etish samaradorligi. Volume 4. CPSU Conference. Toshkent, 2023. -191 – 192 b.
18. Dildor Otajonova. Korpus lingvistikasining shakllanish va rivojlanish tarixidan lavhalar. // Ta’lim jarayonida raqamli texnologiyalarni joriy etish samaradorligi. Volume 4. CPSU Conference 1: pp. 191 – 192.