

KORPUS LINGVISTIKASI VA PARALLEL KORPUSLAR XUSUSIDA

Shakirov Rustam Sobirovich

Buxoro davlat universiteti

Tarjimashunoslik va lingvodidaktika kafedrası o'qituvchisi

ANNOTATSIYA

Ushbu maqolada korpus lingvistikasi, korpus va uning parallel korpus bog'ini haqidagi fikrlar, uning tuzilishi, korpus turlari, token, lemma, stemming haqida so'z boradi. Bugungi kunda korpusning nazariy hamda amaliy ahamiyati o'zbek tilshunosligida tilning mavjud imkoniyatlarini kengroq o'rganishda, tilshunoslikdagi muammoli tomonlarini aniqlash, elektron lug'atlar yaratishda, tilni o'rganishda zamonaviy axborot texnologiyalaridan foydalanish samaradorligini oshirishda, tilda avtomatik tarjima, qidiruv va kompyuter tahlili kabi masalalarni hal qilishda, tilning maxsus sohalar bo'yicha korpusini qurish zarurati mavjudligini ko'rsatadi.

Kalit so'zlar: Korpus, korpus lingvistikasi, parallel korpus, tarjima korpus, taqqoslanadigan korpus, segmentlash, mashina tarjimasi, tokenizatsiya, lemmatizatsiya, stemmizatsiya.

ABSTRACT

This article deals with corpus linguistics, ideas about the corpus and its parallel corpus link, its structure, corpus types, tokens, lemmas, stemming. Today, the theoretical and practical significance of the corpus is in the study of the existing possibilities of language in Uzbek linguistics, the identification of problematic aspects of linguistics, the creation of electronic dictionaries, increasing the effectiveness of modern information technology in language learning, automatic translation, search and computer analysis. In solving problems, there is a need to build a corpus of language in specific areas.

Keywords: Corpus, corpus linguistics, parallel corpus, translation corpus, comparable corpus, segmentation, machine translation, tokenization, lemmatization, stemmization.

KIRISH

Har bir xalq, millat o'z ona tilini asrab-avaylash, saqlash rivojlantirish va albatta uni dunyoga tanitishga harakat qiladi. So'nggi yillarda esa, tilimizga berilayotgan e'tibor yanada o'sdi. Hozirgi kunda o'zbek tilini rivojlantirish, saqlash va kelgusi avlodga bekam-u ko'st yetkazishda axborot-texnologiya vositalaridan foydalanish,

tilimizni modellashtirish, elektron tizimga o'tkazish asosiy, dolzarb mavzulardan biridir. Tilshunoslikning bunday ishlari bilan shug'ullanuvchi yo'nalishi esa, hozirgi kunda zamonaviy tilshunoslik yo'nalishi hisoblangan korpus lingvistikasidir.

Korpus lingvistikasi XX asrning 60-yillarida kompyuter lingvistikasining bo'limi sifatida shakllangan soha bo'lib, kompyuter texnologiyalaridan foydalangan holda lingvistik korpuslarni (matnlar korpuslarini) qurish va ulardan foydalanishning umumiy tamoyillarini ishlab chiqish bilan shug'ullanadi.

Davlatimizda ham bugungi kunda amaliy tilshunosligimiz oldida turgan eng muhim masalalardan biri etib – o'zbek tilining milliy korpusini yaratish siyosat darajasiga ko'tarildi. Xususan, o'zbek tilini jamiyat hayotida hamda xalqaro miqyosda obro'-e'tiborini oshirish maqsadida, o'zbek tiliga oid barcha ilmiy, nazariy va amaliy ma'lumotlarni o'zida jamlagan elektron ko'rinishdagi o'zbek tili milliy korpusini yaratish, o'zbek tilini Internet jahon axborot tarmog'ida ommalashtirish, unda munosib o'rin egallashini ta'minlash, dasturiy mahsulotlarning o'zbekcha ilovalarini yaratish, o'zbek tilini o'rgatuvchi kompyuter dasturlarini keng miqyosda amaliyotga tatbiq qilish, o'zbek tilidagi matnlar tahririga mo'ljallangan kompyuter dasturlarini yaratish ishlari 2 tilshunoslik oldida turgan muhim vazifalar etib belgilab olindi. Amalda bu bo'yicha ko'plab izlanishlar olib borilmoqda. 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.Laurens, E.Etwell, S.Hunston, L.Boizou, McKenneri, J.Grafmiller, J.Grieva, N.Grumbach, S.Hansson, K.McAulif, M.Malberg, P.Milin, A.Murakami, R.Peych, A.Shembri, P.Tompson, B.Vinter, G.Linch kabi xorijiy olimlar¹ tomonidan hamda Turkologiyada korpusshunoslik (korpus lingvistikasi) sohasi bo'yicha ilmiy tadqiqotlar olib borilgan. O'zbek olimlaridan B.Mengliyev, Sh.Shahobiddinova, Z.Xolmanova, S.Karimov, N.Abduraxmonova, L.Raupova, Sh.Hamroyeva, M.Abjalova, G.Toirova, G.Ikromova, J.Djumabayeva, G.Ergasheva, A.Eshmo'minovlar ilmiy ishlar qilishgan. O'zbek tili milliy korpusining konseptologiyasi B.Mengliyev rahbarligidagi olimlar jamoasi tomonidan ishlab chiqilmoqda².

Lingvistik (leksik) korpus yohud til (matnlar) korpusi muayyan lingvistik muammolarni hal qilishga mo'ljallangan, maxsus saralangan va teglangan elektron shakldagi matnlar majmuasi hisoblanadi va pragmatik ahamiyatga ega yirik elektron baza sanaladi. Korpus lingvistikasining predmeti esa til korpuslari sanaladi. Aytish mumkinki, KorpLing o'z predmeti sifatida keng doiradagi foydalanuvchilarning manfaati uchun lingvistik tadqiqotlarga mo'ljallangan til ma'lumotlarining yirik

¹ Abdurakhmonova N. Kompyuter lingvistikasi (darslik) / Globe edit publishing, 2020, 395 b

² Abduraxmonova N. O'zbek tili elektron korpusining kompyuter modellari (monografiya) / Toshkent: Muharrir, 2021, 202 b.

majmualarini yaratish va ulardan foydalanishning nazariy asoslari hamda amaliy mexanizmlariga ega.

E.Finegan darsligida korpus – matn yaratilgan vaziyat va unga tegishli ma'lumotlar, jumladan, so'zlovchi, muallif, adresat yoki auditoriya haqidagi ma'lumotlarni o'z ichiga olgan, mashina o'qiydigan formatga ega reprezentativlangan matnlar majmuasi deya ta'rif beriladi.

T.Mak-Eneri va E.Vilson shunday ta'rif beradi: korpus – til modeli sifatida foydalanish uchun aniq til mezonlari bo'yicha tanlangan til fragment (qism / parcha)lari to'plami.

J.Sinkler – korpus lingvistikasining asoschilaridan biri, korpusni “tilni yoki tilning o'zgarishini to'liq aks ettirish uchun tashqi mezonlarga muvofiq tanlangan elektron shakldagi matnlarning parchalari to'plami” deb ta'riflaydi.

Rus tili milliy korpusida korpus “ma'lum bir tilda elektron shaklda matnlar to'plamiga asoslangan axborot-ma'lumot tizimi” deya ta'riflanadi.

Mazkur ta'riflarda korpusning umumiy va asosiy jihatlari ko'rinadi. Bular: korpusning maqsadli yaratilishi, mashina (maxsus elektron) formatida bo'lishi, reprezentativlik va metalingvistik ma'lumotlarga egalik.

Korpus leksik ma'lumotlar massivi sifatida muayyan davrda tilning leksik tarkibi o'zgarishi jarayonlari dinamikasini o'rganishga, turli janrlarda va turli mualliflarda leksik hamda grammatik xususiyatlarning qo'llanilishini tahlil qilishga imkon beradi. Korpuslar, shuningdek, turli xil tarixiy va zamonaviy lug'atlarni tayyorlashda ko'p qirrali leksikografik ishlarning manbai va vositasi bo'lib xizmat qiladi. Shuningdek, ushbu korpuslardan ta'lim jarayonida tilni o'rgatish uchun grammatikani qurish va takomillashtirishda foydalanish mumkin. Korpus matnlarni statistik tahlil qilish va lingvistik gipotezalarni tekshirish, qoidalarni asoslash yoki yuzaga kelish holatlarini tekshirish uchun foydalaniladigan tizim hisoblanishi bilan ahamiyatli.

O'zbek tili nihoyatda boy va imkoniyatlari keng til sanaladi. Leksik qatlamning xilma-xilligi, jozibadorligi, o'ziga xos iboralar, o'xshatish, ko'chimlar, sintaktik figuralar badiiy asarlarda keng tasvirlangan. Agar bularning hammasi tadqiq qilinib, o'zbek tilining milliy korpusi yaratilsa, kompyuter tizimiga o'tkazilib elektronlashtirilsa, tilshunosligimizda amalga oshirilgan juda katta yutuq bo'lardi. Rivojlangan korpus lingvistikasi tilshunoslikning barcha sohalarini qamrab olishi mumkin. Misol tariqasida, dialektologiya sohasini olaylik. Bu sohada kompyuter lingvistikasining o'rni beqiyosdir. Bebaho xazinamiz sanalgan shevalar yig'ilib, milliy korpus talablari asosida tizimlashtirilsa, o'chmas bo'lib muhrlangan bo'lardi. Har bir shevaning qaysi hududda nimani nomlashi, adabiy tilda qaysi so'zga teng kelishi va hokazolar bilan bir buyruq amali orqali juda qulay tarzda tanishish imkonini beradi.

Umuman, Korpus Lingvistikasi XXI asr ilmi hisoblanadi. U bugungi kunda jahon miqyosida lingvistikaning “tanasi”, kompyuter lingvistikasining esa “dvigateli” sanaladi. Bugungi kunda korpus lingvistikasi tilshunoslikda nisbatan yangi yondashuv sifatida tushuniladi, bu vaziyat tilni “real hayotda” kompyuterlar va elektron korpuslar orqali ishlatishni o‘rganish bilan bog‘lanadi.

XULOSA.

1. Til korpusi – milliy til birliklarining xususiyatlarini aniqlash maqsadida elektron qidiruv imkoniyati mavjud tizim, tabiiy tilning raqamlashgan yozma va og‘zaki matnlari jamlanmasi.

2. Undan lingvistlar, leksikograflar, kompyuter lingvistlari, dasturchilar, muharrirlar, tarjimonlar, jurnalistlar, noshirlar, olimlar, o‘qituvchilar, ta’lim oluvchilar va boshqa har qanday soha mutaxassisi keng foydalanish imkoniyatiga ega bo‘ladi.

3. Til ta’limi hayotdan uzilib qolmasligi uchun muntazam ravishda matnlar bazasi yangilanib boriladigan korpusga tayanish maqsadga muvofiq bo‘ladi.

REFERENCES:

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

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

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