O‘ZBEK TILINING TA’LIMIY KORPUSINI LOYIHALASH: LINGVODIDAKTIK TAMOYILLAR VA ANNOTATSIYA MODELI

Saatova Surayyo
O‘zbekiston Milliy universiteti
dotsent v.b., PhD.
ORCID: 0009-0005-7251-6684

https://doi.org/10.5281/zenodo.24258434

Annotatsiya. *Maqolada o‘zbek tilining ta’limiy korpusini loyihalash tamoyillari va annotatsiya modeli yoritiladi. Ta’limiy korpus uchun darslik matnlari, o‘quv lug‘atlari, adaptatsiyalangan badiiy matnlar, sinf nutqi va o‘rganuvchi yozma ishlarini qamrab oluvchi besh subkorpusli model taklif etiladi. Shuningdek, metama’lumotlar, morfologik, sintaktik va xatolik annotatsiyasidan iborat to‘rt qatlamli model asoslanadi. O‘zbek tilining agglutinativ xususiyatlari hamda grafik variantlarini me’yorlashtirish masalalari ko‘rib chiqiladi.*

Kalit so‘zlar: *ta’limiy korpus, korpus lingvistikasi, lingvodidaktika, o‘rganuvchi korpusi, annotatsiya, morfologik tahlil.*

Abstract. *The article discusses the design principles and annotation model of an educational corpus of the Uzbek language. A five-subcorpus model comprising textbook texts, educational dictionaries, adapted literary texts, classroom discourse, and learner writing is proposed. A four-layer annotation model covering metadata, morphological, syntactic, and error annotation is presented. Particular attention is given to the agglutinative nature of Uzbek and script normalization.*

Keywords: *educational corpus, corpus linguistics, linguodidactics, learner corpus, annotation, morphological analysis.*

Аннотация. *В статье рассматриваются принципы проектирования учебного корпуса узбекского языка и модель его аннотирования. Предлагается пятикомпонентная модель, включающая тексты учебников, учебные словари, адаптированные художественные тексты, аудиторную речь и письменные работы обучающихся.*

Обосновывается четырехслойная модель аннотации: метаданные, морфологическая, синтаксическая и ошибочная аннотация. Особое внимание уделяется агглютинативной природе узбекского языка и нормализации графических вариантов.

Ключевые слова: *учебный корпус, корпусная лингвистика, лингводидактика, корпус обучающихся, аннотация, морфологический анализ.*

Kirish. Korpus lingvistikasi til birliklarining real qo‘llanishini elektron matnlar asosida o‘rganish imkonini beruvchi zamonaviy metodologik yo‘nalishlardan biridir. Korpusga asoslangan tahlil til hodisalarini faqat nazariy tavsif yoki tadqiqotchining intuitiv kuzatuvlari orqali emas, balki muayyan lingvistik birliklarning qo‘llanish chastotasi, konteksti va o‘zaro munosabatlari asosida o‘rganishga imkon beradi (Sinclair, 1991; McEnery, Hardie, 2012).

Ushbu yondashuvning til ta’limiga tatbiqi o‘quv materiallarini tanlash, leksik minimumni belgilash, grammatik birliklarning qo‘llanishini o‘rganish hamda o‘rganuvchilar nutqini tahlil qilish uchun yangi imkoniyatlar yaratadi.

Korpusdan ta’lim jarayonida foydalanishning muhim yo‘nalishlaridan biri Data-Driven Learning (DDL) yondashuvidir. Mazkur yondashuvda o‘rganuvchi tayyor qoidalarni o‘zlashtirish bilan cheklanmay, korpusdagi real misollarni kuzatish, qiyoslash va tahlil qilish orqali til qonuniyatlarini mustaqil aniqlaydi (Johns, 1991).

O‘rganuvchi korpuslari esa til o‘zlashtirish jarayonida yuzaga keladigan xususiyatlar va xatoliklarni empirik o‘rganishga xizmat qiladi (Granger, 1998).

O‘zbek tilining raqamli infratuzilmasini rivojlantirish doirasida milliy korpus va avtomatik lingvistik tahlil vositalarini shakllantirishga e’tibor kuchaymoqda. O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi PF–5850-son Farmonida o‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini oshirish bilan bir qatorda, uning zamonaviy axborot-kommunikatsiya texnologiyalaridagi imkoniyatlarini kengaytirish masalalari ham belgilangan.

O‘zbek tilining milliy korpusi va ta’limiy korpusini rivojlantirish ushbu vazifalar bilan uzviy bog‘liq.

Biroq ta’limiy korpusni milliy korpusning hajman kichraytirilgan ko‘rinishi sifatida talqin qilish uning lingvodidaktik vazifalarini to‘liq namoyon etmaydi. Ta’limiy korpus materiallarni o‘quv maqsadlari, til o‘zlashtirish bosqichlari va foydalanuvchi ehtiyojlariga muvofiq tanlashni talab qiladi. Shu nuqtayi nazardan, o‘zbek tilining ta’limiy korpusini loyihalashda uning tarkibi, annotatsiya tizimi va foydalanish mexanizmlarini yagona konseptual model asosida belgilash dolzarb hisoblanadi.

Tadqiqotning maqsadi o‘zbek tilining ta’limiy korpusini shakllantirishga xizmat qiluvchi lingvodidaktik tamoyillarni aniqlash, uning subkorpuslar tarkibini asoslash hamda lingvistik va o‘quv ma’lumotlarini birlashtiruvchi annotatsiya modelini taklif etishdan iborat.

Tadqiqot metodologiyasi. Tadqiqotda korpus lingvistikasi, lingvodidaktika va o‘rganuvchi korpusi nazariyalariga asoslangan tavsifiy-tahliliy yondashuv qo‘llanildi. Xalqaro korpus tadqiqotlarida ishlab chiqilgan reprezentativlik, muvozanatlilik va tanlanma tamoyillari ta’limiy korpusning maqsad hamda vazifalari bilan bog‘liq holda ko‘rib chiqildi (Biber, Conrad, Reppen, 1998; McEnery, Hardie, 2012).

Modelni ishlab chiqishda Design-Based Research (DBR) metodologiyasining loyihalashga yo‘naltirilgan tamoyillaridan foydalanildi. Mazkur yondashuv nazariy asoslar va amaliy ehtiyojlar o‘rtasidagi bog‘liqlikni ta’minlash, taklif etilayotgan modelni keyingi pilot sinovlar orqali takomillashtirish imkonini beradi. Tadqiqot konseptual xarakterga ega bo‘lib, unda ishlab chiqilgan modelning to‘liq hajmdagi korpusda empirik sinovdan o‘tkazilishi kelgusi bosqich vazifasi sifatida belgilanadi.

Ta’limiy korpusni loyihalash tamoyillari. Ta’limiy korpusning samaradorligi, avvalo, uning tarkibidagi materiallarning o‘quv maqsadlariga muvofiqligi bilan belgilanadi. Shu sababli korpusni shakllantirishda quyidagi tamoyillarga tayanish maqsadga muvofiq.

Reprezentativlik korpus materiallarining o‘rganilayotgan til hodisalari va ta’limiy vazifalarni yetarli darajada aks ettirishini nazarda tutadi. Ta’limiy korpusda bu tamoyil turli janr, mavzu, til birliklari va o‘quv bosqichlariga tegishli materiallarni qamrab olish orqali amalga oshiriladi.

Muvozanatlilik korpus tarkibidagi materiallarning belgilangan mezonlar bo‘yicha nisbatini ta’minlaydi. Subkorpuslar ulushi ta’limiy maqsad, o‘quv dasturi va foydalanuvchi ehtiyojlariga muvofiq belgilanadi.

Lingvodidaktik yo‘naltirilganlik materiallarni o‘quvchining yoshi, til bilish darajasi va ta’lim bosqichiga mos tanlashni talab qiladi. Bu tamoyil korpusdan leksik minimum, grammatik ketma-ketlik va matn murakkabligini aniqlashda foydalanish imkoniyatini kengaytiradi.

Kengaytiriluvchanlik korpusga yangi materiallar va annotatsiya turlarini bosqichma-bosqich qo‘shish, uning tarkibini yangilash hamda texnologik imkoniyatlarini rivojlantirishni nazarda tutadi.

O‘zbek tilining ta’limiy korpusi: konseptual model

Ta’limiy korpusning asosiy xususiyati unda til ta’limi jarayonida foydalaniladigan materiallar bilan o‘rganuvchilar yaratgan nutqiy mahsulotlarning bir tizimda birlashtirilishidir.

Shu asosda o‘zbek tilining ta’limiy korpusi uchun quyidagi besh subkorpusdan iborat konseptual model taklif etiladi.

Subkorpus	Asosiy manbalar	Taklif etilayotgan ulush
Darslik matnlari	Ona tili, adabiyot va boshqa fan darsliklari	35 %
O‘quv lug‘atlari	Izohli, sinonim, terminologik va imlo lug‘atlari	15 %
Adaptatsiyalangan badiiy matnlar	O‘quvchilar yoshiga moslashtirilgan badiiy asarlar	20 %
Sinf nutqi	O‘qituvchi va o‘quvchilarning dars jarayonidagi nutqi	10 %
O‘rganuvchi korpusi	Insho, bayon va boshqa yozma ishlar	20 %

Izoh: foizlar korpusning yakuniy empirik tarkibini emas, loyihalash bosqichida taklif etilayotgan dastlabki nisbatlarni bildiradi.

Modelning muhim jihati o‘rganuvchi korpusining ta’limiy korpus tarkibiga integratsiyalanishidir. An’anaviy o‘quv materiallari o‘rganuvchiga taqdim etiladigan til birliklarini aks ettirsa, o‘rganuvchi korpusi uning amaliy nutqida namoyon bo‘ladigan leksik, grammatik va imloviy xususiyatlarni qayd etadi. Mazkur ikki yo‘nalishning birlashtirilishi taqdim etilayotgan til materiali bilan uning o‘zlashtirilish natijalarini qiyosiy o‘rganish imkonini beradi.

Korpus tarkibidagi materiallarni o‘quv bosqichlari, mavzular, janrlar va til bilish darajalari bo‘yicha metama’lumotlar bilan ta’minlash ularning izlanish va tahlil jarayonida samarali qo‘llanishini ta’minlaydi.

Annotatsiya modeli va o‘zbek tilining lingvistik xususiyatlari. Ta’limiy korpusning lingvistik imkoniyatlari undagi materiallarning annotatsiya qilinish darajasiga bog‘liq.

Annotatsiya matn tarkibidagi birliklarni belgilangan lingvistik va texnik mezonlar asosida tavsiflash hamda ularni avtomatik yoki qo‘lda tahlil qilish imkonini beradi. Tadqiqotda ta’limiy korpus uchun to‘rt qatlamli annotatsiya modeli taklif etiladi.

Annotatsiya qatlami	Asosiy mazmuni
Metama’lumotlar	Matn manbasi, muallif, sana, janr, mavzu, ta’lim bosqichi, til darajasi
Morfologik annotatsiya	So‘z shakli, lemma, so‘z turkumi, grammatik kategoriyalar
Sintaktik annotatsiya	Gap tuzilishi, sintaktik munosabatlar, bog‘lanishlar
Xatolik annotatsiyasi	Imlo, leksik, morfologik, sintaktik va punktuatsion xatolar

Metama’lumotlar korpus materiallarini saralash va ularning kelib chiqishini aniqlashga xizmat qiladi. Morfologik annotatsiya so‘z shakllarini lemma va grammatik kategoriyalar bilan bog‘lash imkonini beradi. Sintaktik annotatsiya esa gap tarkibidagi birliklar o‘rtasidagi munosabatlarni tavsiflaydi. Xatolik annotatsiyasi o‘rganuvchi korpusidagi tipik xatolarni tasniflash va ularning takrorlanish xususiyatlarini o‘rganishga yo‘naltiriladi.

Annotatsiya tizimini ishlab chiqishda TEI P5 va Universal Dependencies (UD) standartlaridan foydalanish ma’lumotlarning strukturaviy izchilligini ta’minlash hamda boshqa korpuslar bilan integratsiya imkoniyatlarini kengaytirishga xizmat qiladi (TEI Consortium; Nivre et al., 2020). Bunda xalqaro standartlarni o‘zbek tilining grammatik xususiyatlariga moslashtirish muhim metodologik vazifa hisoblanadi.

O‘zbek tilining agglutinativ tuzilishi morfologik annotatsiyada alohida e’tiborni talab qiladi. So‘z shakllarining ko‘p morfemalardan tashkil topishi, grammatik omonimiya va ayrim yordamchi fe’lli konstruksiyalar avtomatik tahlil jarayonida murakkabliklar yuzaga keltirishi mumkin. Shu sababli lemma, morfemik tarkib va grammatik belgilarning izchil tavsiflanishi korpusning lingvistik aniqligi uchun muhimdir.

Lotin va kirill grafikalarining parallel qo‘llanishi ham ta’limiy korpusni shakllantirishda hisobga olinishi lozim. Apostrof belgilarining turli kodlash variantlari, grafik shakllarning farqlanishi va imloviy me’yorlarning o‘zgarishi qidiruv hamda statistik tahlil natijalariga ta’sir ko‘rsatishi mumkin. Grafik me’yorlashtirish jarayonida matnning asl shaklini saqlash va normallashtirilgan variantini alohida qayd etish maqsadga muvofiq.

Ta’limiy korpusning lingvodidaktik imkoniyatlari. Taklif etilayotgan model ta’lim jarayonida korpusga asoslangan bir qator amaliy vazifalarni bajarish imkonini beradi. Birinchidan, turli o‘quv bosqichlarida qo‘llanadigan leksik birliklarning chastotasi va kontekstlarini tahlil qilish orqali leksik minimumni empirik ma’lumotlar asosida shakllantirish mumkin (Nation, 2001; Coxhead, 2000).

Ikkinchidan, darslik va adaptatsiyalangan matnlardagi leksik hamda grammatik birliklarni qiyoslash matn murakkabligini baholash mezonlarini ishlab chiqishga zamin yaratadi. Bunday tahlilda leksik chastota, so‘zlarning takrorlanishi, grammatik konstruksiyalar va gap uzunligi kabi ko‘rsatkichlardan foydalanish mumkin. Biroq ularning o‘zbek tili uchun samaradorligi maxsus empirik sinovlar orqali aniqlanishi lozim.

Uchinchidan, o‘rganuvchi korpusi asosida yozma nutqdagi tipik xatolarni aniqlash, ularning darajalar va ta’lim bosqichlari bo‘yicha taqsimotini o‘rganish imkoniyati yuzaga keladi.

Bu ma’lumotlar o‘quv materiallarini takomillashtirish va diagnostik topshiriqlarni ishlab chiqishda qo‘llanishi mumkin.

To‘rtinchidan, korpus konkordanslari, kollokatsiyalar va real kontekstlardan foydalanish DDL yondashuvi asosida mustaqil o‘rganish topshiriqlarini yaratishga imkon beradi. Bunda o‘quvchi til birliklarining qo‘llanishini kuzatish, ularni qiyoslash va lingvistik xulosalar chiqarish faoliyatiga jalb etiladi (O’Keeffe, McCarthy, Carter, 2007).

Xulosa. O‘zbek tilining ta’limiy korpusini loyihalashda uning lingvodidaktik maqsadi, tarkibiy muvozanati va annotatsiya tizimini o‘zaro bog‘liq holda belgilash muhim hisoblanadi.

Ta’limiy korpus milliy korpusning ixtisoslashgan ko‘rinishi bo‘lish bilan birga, o‘quv materiallari va o‘rganuvchilar nutqini yagona empirik tizimda birlashtiruvchi mustaqil tadqiqot vositasi sifatida namoyon bo‘ladi.

Taklif etilgan model o‘zbek tilining ta’limiy korpusini shakllantirish, leksik minimumni aniqlash, o‘quv matnlari murakkabligini baholash va o‘rganuvchilar nutqini tahlil qilish uchun metodologik asos bo‘lib xizmat qilishi mumkin.

Modelning amaliy samaradorligini aniqlash uchun kelgusida pilot korpus yaratish, annotatsiya aniqligini baholash hamda subkorpuslar nisbatini empirik ma’lumotlar asosida takomillashtirish zarur.

Foydalanilgan adabiyotlar

1. Biber D., Conrad S., Reppen R. *Corpus Linguistics: Investigating Language Structure and Use*. – Cambridge: Cambridge University Press, 1998.
2. Coxhead A. A New Academic Word List // *TESOL Quarterly*. – 2000. – Vol. 34, No. 2. – P. 213–238.
3. Granger S. (ed.). *Learner English on Computer*. – London; New York: Longman, 1998.
4. Johns T. Should You Be Persuaded: Two Samples of Data-Driven Learning Materials // *ELR Journal*. – 1991. – Vol. 4. – P. 1–16.
5. McEnery T., Hardie A. *Corpus Linguistics: Method, Theory and Practice*. – Cambridge: Cambridge University Press, 2012.
6. Nivre J., de Marneffe M.-C., Ginter F. et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // *Proceedings of LREC 2020*. – Marseille, 2020. – P. 4034–4043.
7. Toirova G. *O‘zbek tili milliy korpusini yaratishning nazariy va amaliy masalalari*: filol. fanlari bo‘yicha falsafa doktori (PhD) diss. avtoreferati. – Buxoro, 2021.
8. Elov B.B., Sirojiddinov Sh., Hamroyeva Sh. va b. O‘zbek tili matnlarini avtomatik teglash masalalari // *Kompyuter lingvistikasi: muammolar, yechim, istiqbollar*: xalqaro ilmiy-amaliy konferensiya materiallari. – Toshkent: TDO‘TAU, 2023.
9. O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi PF–5850-son Farmoni “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”. – Toshkent, 2019.
10. O‘zbek tilining milliy korpusi. URL: https://uzbekcorpus.uz
11. O‘zbek tilining ta’limiy korpusi. URL: https://uzschoolcorpara.uz