O‘ZBEK TILINI IKKINCHI TIL SIFATIDA O‘RGANUVCHILARNING YOZMA NUTQI KORPUSI AVTOMATLASHTIRILGAN LINGVISTIK TAHLIL ASOSI SIFATIDA

Solijonova Mohinur Asrorjon qizi
Magistrant, Farg‘ona davlat universiteti
Tel.: +998 90 293 30 16
E-mail: solijonovam2002@gmail.com
https://doi.org/10.5281/zenodo.22077445

Annotatsiya: Mazkur maqolada o‘zbek tilini ikkinchi til sifatida o‘rganuvchilarning yozma nutqi korpusini shakllantirish va undan avtomatlashtirilgan tahlilda foydalanish masalalari yoritiladi. Korpusning metama’lumotlar, lingvistik va xato annotatsiyasi bilan ta’minlanishi hamda o‘zbek tilining agglutinativ xususiyatlarini hisobga olish zarurati asoslanadi. Korpusning xatolarni aniqlash va yozma nutqni avtomatik baholash uchun empirik baza sifatidagi imkoniyatlari ko‘rsatiladi.
Kalit so‘zlar: o‘zbek tili, ikkinchi til, yozma nutq, o‘rganuvchi korpusi, xatolar tahlili, avtomatlashtirilgan tahlil, NLP, avtomatik baholash.

Аннотация: В статье рассматриваются вопросы формирования корпуса письменной речи изучающих узбекский язык как второй язык и его использования для автоматизированного анализа. Обосновывается необходимость включения метаданных, лингвистической и ошибочной аннотации, а также учета агглютинативных особенностей узбекского языка. Показаны возможности корпуса как эмпирической базы для выявления ошибок и автоматического оценивания письменной речи.
Ключевые слова: узбекский язык, второй язык, письменная речь, корпус учащихся, анализ ошибок, автоматизированный анализ, NLP, автоматическое оценивание.

Abstract: This article examines the development of a corpus of written language produced by learners of Uzbek as a second language and its application in automated analysis. The study substantiates the need to include metadata, linguistic and error annotation, as well as to account for the agglutinative features of Uzbek. The corpus is considered as an empirical basis for error detection and automated assessment of written language.
Keywords: Uzbek language, second language, written language, learner corpus, error analysis, automated analysis, NLP, automated assessment.

Raqamli texnologiyalarning tilshunoslik va ta’limga joriy etilishi ikkinchi til o‘rganuvchilarning yozma nutqini real ma’lumotlar asosida tadqiq etish imkonini kengaytirmoqda. Korpus texnologiyalari orqali o‘rganuvchilarning yozma mahsulotlarini tizimli yig‘ish, lingvistik annotatsiyalash va statistik tahlil qilish, shu asosda tilni o‘zlashtirish qonuniyatlarini aniqlash mumkin.

O‘rganuvchi yozma nutqi korpusi umumiy til korpusidan farqli ravishda nafaqat me’yoriy birliklarni, balki o‘rganuvchilar xatolarini ham tadqiqot materiali sifatida qamrab oladi. Bunday ma’lumotlar morfologik, leksik va sintaktik birliklarning o‘zlashtirilish darajasini aniqlash, yozma kompetensiyani baholash va avtomatlashtirilgan lingvistik tizimlarni ishlab chiqishga xizmat qiladi.

O‘zbek tili uchun Universal Dependencies asosidagi morfologik va sintaktik resurslar shakllanmoqda [1; 2; 4]. Biroq mavjud resurslar ikkinchi til o‘rganuvchilarining xatolari va yozma kompetensiyasini to‘liq aks ettirmaydi. Shu bois maxsus o‘rganuvchi korpusini yaratish dolzarb bo‘lib, maqolaning maqsadi uning shakllantirish tamoyillari va avtomatlashtirilgan tahlildagi ahamiyatini asoslashdan iborat.

O‘rganuvchi korpusi muayyan tilni o‘rganayotgan shaxslar tomonidan yaratilgan, metama’lumotlar va lingvistik annotatsiya bilan ta’minlangan yozma matnlarning tizimli majmuasidir. Uning asosiy xususiyati tilning me’yoriy holatini emas, balki o‘zlashtirish jarayonini aks ettirishidir. Shu sababli korpus lingvodidaktika, xatolar tahlili va kompyuter lingvistikasi uchun muhim empirik manba hisoblanadi. O‘rganuvchi xatolari oddiy kamchilik emas, balki til tizimining o‘zlashtirilish darajasini ko‘rsatuvchi lingvistik ma’lumot sifatida qaraladi. Ularning chastotasi va taqsimotini tahlil qilish turli darajadagi o‘rganuvchilarning yozma nutqi o‘rtasidagi farqlarni aniqlash imkonini beradi. Korpusning reprezentativligini ta’minlashda o‘rganuvchining til darajasi, topshiriq turi, mavzu va matn hajmi hisobga olinishi, turli kommunikativ vazifalarni aks ettiruvchi materiallar kiritilishi lozim. Matnlar bilan bog‘liq metama’lumotlarni standart shaklda saqlash ularni guruhlash va natijalarni taqqoslashga, original matnni saqlash esa avtomatik xato aniqlash va keyingi tahlilga imkon beradi. Ekspert tomonidan tuzatilgan variant mavjud bo‘lsa, uni original matndan alohida saqlash maqsadga muvofiq.

Lingvistik annotatsiya matnni tokenlarga ajratish, lemmatizatsiya, so‘z turkumlari, morfologik xususiyatlar va sintaktik bog‘lanishlarni belgilashni qamrab oladi. O‘zbek tili uchun mavjud Universal Dependencies resurslari mazkur jarayonni tashkil etishda metodologik asos bo‘la oladi [1; 2; 4]. Biroq o‘rganuvchi korpusida umumiy lingvistik annotatsiya bilan bir qatorda xato annotatsiyasini ham joriy etish zarur. Unda xatoli birlik, uning turi va tuzatilgan varianti belgilanib, xatolarni tizimli tahlil qilish hamda avtomatik aniqlash algoritmlarini ishlab chiqish uchun asos yaratiladi. O‘zbek tilining agglutinativ tuzilishi esa alohida metodologik talablarni yuzaga keltiradi: grammatik ma’nolar affikslar ketma-ketligi orqali ifodalangani sababli xato butun so‘zga emas, muayyan morfema yoki grammatik ko‘rsatkichga tegishli bo‘lishi mumkin. Masalan, o‘rganuvchi so‘z asosini to‘g‘ri qo‘llab, unga noto‘g‘ri kelishik yoki egalik qo‘shimchasini biriktirishi mumkin. Shu bois annotatsiyada morfemik tarkib va grammatik kategoriyalarni hisobga olish o‘zbek tilini ikkinchi til sifatida o‘zlashtirishdagi tipik qiyinchiliklarni aniqlash imkonini beradi.

Korpusga o‘rganuvchining til darajasi bilan birga ekspert baholarini kiritish ham muhimdir. Yozma ishning umumiy bahosi va asosiy lingvistik mezonlar bo‘yicha natijalar lingvistik belgilar bilan baho o‘rtasidagi bog‘liqlikni aniqlashga xizmat qiladi. Natijada korpus nafaqat o‘rganuvchi xatolarini o‘rganish vositasi, balki yozma nutq sifatini miqdoriy tavsiflash uchun empirik baza vazifasini ham bajaradi.

O‘rganuvchi korpusining amaliy qiymati uning avtomatlashtirilgan lingvistik tahlil uchun qo‘llanish imkoniyati bilan belgilanadi. Bunday tahlil tokenizatsiya, lemmatizatsiya, morfologik va sintaktik tahlil hamda xatolarni aniqlash kabi jarayonlarni qamrab oladi. Natijada matndan leksik xilma-xillik, gaplarning o‘rtacha uzunligi, grammatik xatolar zichligi va sintaktik murakkablik kabi miqdoriy belgilar ajratib olinadi. Ularni matn hajmiga nisbatan hisoblash turli hajmdagi yozma ishlarni obyektiv taqqoslash imkonini beradi.

Ajratilgan lingvistik belgilarni o‘rganuvchining til darajasi va ekspert baholari bilan bog‘lash orqali yozma kompetensiyaning rivojlanish dinamikasi hamda avtomatik baholash uchun ahamiyatli ko‘rsatkichlarni aniqlash mumkin. Bunda faqat xatolar soniga tayanish yetarli emas. Yozma nutq sifati grammatik aniqlik bilan birga leksik imkoniyat, sintaktik murakkablik, mazmuniy izchillik va kommunikativ vazifaning bajarilishi kabi xususiyatlar

bilan ham belgilanadi. Shuning uchun korpusdan olinadigan lingvistik belgilar yozma kompetensiyaning bir nechta jihatlarini qamrab olishi lozim.

Korpusning tashkil etilish sifati kelajakdagi avtomatik baholash modelining ishonchliligiga bevosita ta’sir qiladi. Ma’lumotlarni o‘quv va test to‘plamlariga ajratishda bir o‘rganuvchining matnlarini turli to‘plamlarga tasodifiy kiritish modelning individual yozish uslubiga moslashishiga olib kelishi mumkin. Shu sababli ma’lumotlarni o‘rganuvchilar kesimida ajratish va modelni ilgari uchramagan matnlarda sinash metodologik jihatdan to‘g‘ri hisoblanadi.

Korpusni dastlab pilot shaklda yaratish, annotatsiya kategoriyalarini sinash va keyinchalik ma’lumotlar hajmini kengaytirish maqsadga muvofiq. Yetarli hajmdagi va sifatli annotatsiyalangan ma’lumotlar bazasi shakllangach, undan avtomatik xato aniqlash hamda yozma nutqni baholash modellarini ishlab chiqishda foydalanish mumkin. Shu nuqtai nazardan, o‘rganuvchi korpusi dissertatsiyada ishlab chiqiladigan avtomatik baholash modelining lingvistik va empirik asosini tashkil etadi.

O‘zbek tilini ikkinchi til sifatida o‘rganuvchilarning yozma nutqi korpusini yaratish o‘rganuvchilarning real yozma mahsulotlarini tizimli tadqiq etish, xatolarni aniqlash va yozma kompetensiyani obyektiv baholash uchun muhim ilmiy asos yaratadi. Korpusda original matnlar, metama’lumotlar, lingvistik va xato annotatsiyasi hamda ekspert baholarining saqlanishi uning tadqiqot va amaliy qiymatini oshiradi. O‘zbek tilining agglutinativ xususiyati esa morfemik tarkib va grammatik ko‘rsatkichlarni hisobga oluvchi annotatsiya tizimini talab qiladi.

Mavjud o‘zbek Universal Dependencies resurslari avtomatik morfologik va sintaktik tahlil uchun muhim metodologik baza yaratgan bo‘lsa-da, ikkinchi til o‘rganuvchilarining yozma nutqi va xatolarini maxsus aks ettiruvchi korpusga ehtiyoj mavjud. Bunday korpus avtomatik lingvistik tahlil uchun ma’lumotlar bazasi bo‘lishi bilan birga, o‘zbek tilini ikkinchi til sifatida o‘rganuvchilarning yozma nutqini avtomatik baholash modelini yaratish uchun zarur empirik manba vazifasini bajaradi.

Foydalanilgan adabiyotlar

1. Matlatipov S. G., Aripov M. UzUDT: Uzbek Universal Dependencies Treebank. 2026.
2. Matlatipov S. G., Aripov M., Bobokandov M., Matlatipov G. Annotated Universal Dependencies Dataset for Literary and Educational Uzbek Texts. Data in Brief, 2026, Vol. 66.
3. Normamatova M., Abduraxmonova N. O‘zbek tilini ikkinchi til sifatida o‘rganish jarayonida til korpusining roli va samaradorligi. 2025.
4. Akhundjanova A., Talamo L. Universal Dependencies Treebank for Uzbek. 2025.
5. Council of Europe. Common European Framework of Reference for Languages: Companion Volume. Strasbourg, 2020.
6. Granger S. Learner English on Computer. London: Longman, 1998.