

O'ZBEK AKADEMIK MATNLARIDA NOMUVOFIQLIKLARNI KO'P QATLAMLI ANNOTATSIYALASH SXEMASI

Daminova Faridaxon Bahriddin qizi

daminovafaridaxon@gmail.com

Guliston shahar 6-maktab o'qituvchisi

Annotatsiya Mazkur maqolada o'zbek akademik yozuv korpusidagi nomuvofiqliklarni belgilash uchun ko'p qatlamli annotatsiyalash sxemasi taklif etiladi. Sxema uchta asosiy funksional sinf – imloviy-grafik (IML), funksional-uslubiy (USL) va terminologik (TER) – asosida qurilgan bo'lib, ularning har biri o'ziga xos ichki teglar, dalil manbai va tahrir maqomi bilan ta'minlanadi. Annotatsiya birligi nafaqat til shaklini, balki me'yoriy asos, kontekst, tavsiyaning majburiylik darajasi va semantik xavfni ham qamrab oladi. Maqolada yettita sathdan iborat ma'lumot modeli, annotatsiya yozuvining asosiy maydonlari, span chegaralash konvensiyalari hamda annotatsiyalash yo'riqnomasi va qaror daraxti tavsiflanadi. Taklif etilayotgan sxema keyingi avtomatik model uchun ishonchli lingvistik referens qatlamni shakllantirishga xizmat qiladi.

Kalit so'zlar: *ko'p qatlamli annotatsiyalash, o'zbek akademik korpusi, imloviy-grafik tahlil (IML), funksional-uslubiy tahlil (USL), terminologik annotatsiya (TER), span annotatsiyasi, annotatsiya sxemasi.*

Kirish. O'zbek akademik yozuv korpusiga nomuvofiqliklar haqida lingvistik axborot kiritish oddiy “xato – to'g'ri” juftliklarini yig'ish bilan chegaralanmaydi. Ilmiy matnda muayyan shakl rasmiy imlo me'yorini buzishi, boshqa shakl janrga nomuvofiq bo'lishi yoki bir termin turli ma'nolarda qo'llanishi mumkin. Bu hodisalarning barchasini bitta tahrir yorlig'iga birlashtirish modelga noto'g'ri nazorat signali beradi: imlo qoidasi buzilishi zaruriy tuzatishni, janriy maqbullik esa ko'pincha muallif bilan kelishiladigan tavsiyani talab qiladi. Shuning uchun annotatsiya birligi til shakli bilan birga me'yoriy asos, kontekst, tavsiyaning majburiylik darajasi va semantik xavfni ham qayd etishi lozim.

Taklif etilayotgan annotatsiyalash qatlami hujjat, janr, fan sohasi va matn holati bo'yicha tashkil etilgan korpus tuzilmasi ustiga quriladi. Tahlilning tayanch funksional sinflari uchtdan iborat: IML (imloviy-grafik), USL (funksional-uslubiy) va TER (terminologik). Bu sinflar o'zaro mustaqil emas – bitta birlik imloviy shakl jihatidan noto'g'ri va ayni paytda termin sifatida soha bilan bog'liq bo'lishi mumkin. Bunday holatda takroriy xato yaratmasdan, bir hodisaning turli talqinlarini bog'lovchi mexanizm zarur bo'ladi. Ushbu maqola aynan shu mexanizmni –

annotatsiya birligi, sathlar ierarxiyasi, funksional sinflar va qaror qabul qilish tartibini – tizimli tarzda bayon etadi.

Annotatsiya birligi, kontekst ko‘lami va span konvetsiyasi

Korpusda hujjat – kelib chiqish va statistik umumlashtirishning asosiy birligi; nomuvofiqlik esa odatda hujjat ichidagi muayyan matn oralig‘i yoki ikki birlik orasidagi pozitsiya orqali ifodalanadi. Hujjat → bo‘lim → paragraf → gap → token → belgi ierarxiyasi annotatsiyaga izchil manzil beradi. Biroq har bir nomuvofiqlikni token chegarasiga majburan sig‘dirish to‘g‘ri emas: so‘z birikmasining ortiqcha takrorlanishi, referentning noaniqligi yoki butun paragrafdagi sabab–oqibat uzilishi kattaroq kontekstni talab qiladi. Shu bois “annotatsiya joyi” va “qarorni tekshirish uchun zarur kontekst” ikki alohida atribut sifatida yuritiladi.

Span boshlanishi va oxiri korpusning versiyalangan, o‘zgartirilmaydigan matn tasvirida belgilanadi. Taklif etilayotgan texnik konvetsiyada chegaralar Unicode kod nuqtalari indeksleri bo‘yicha yarim ochiq [b, e) interval sifatida yoziladi: boshlanish indeksi intervalga kiradi, tugash indeksi esa kirmaydi. Nol uzunlikdagi [b, b) oralig‘i tushib qolgan tinish belgisi singari kiritish amalini ko‘rsatishi mumkin. Dasturiy muhitda UTF-8 bayt ofseti, UTF-16 kod birliklari va foydalanuvchi ko‘radigan grafema chegaralari har doim bir xil bo‘lmagani sababli, eksportda tanlangan indekslash birligi albatta ko‘rsatiladi.

Ko‘p qatlamli model: annotatsiyaning yettita sathi

Taklif etilayotgan ma’lumot modeli annotatsiyani bitta tekis yorliqlar ro‘yxati sifatida emas, balki yettita bog‘liq sathdan iborat tuzilma sifatida ifodalaydi – manba hujjatidan tortib, audit izigacha. Har bir sath o‘z tekshiruv savoliga javob beradi va ustki sathlar uchun tayanch vazifasini bajaradi.

1-jadval. Ko‘p qatlam model sathilari

№	Sath	Annotatsiya obyekti	Asosiy belgilar	Tekshiruv savoli
1.	Manba	Hujjat va versiya	document_id, source_hash, genre, domain	Qaysi material va qaysi nusxa?
2.	Segment	Gap, token, span, pozitsiya	sentence_id, start, end, unit_scope	Nomuvofiqlik qayerda joylashgan?
3.	Tasnif	Funksional va ichki sinf	primary_type, subtype, linguistic_level	Qaysi lisoniy hodisa?
4.	Hukm	Me'yoriy maqom va xavf	decision_status, semantic_risk	Majburiy tuzatishmi yoki tavsiyasi?
5.	Dalil	Qoida, kontekst, termin manbasi	evidence_id, norm_scope, confidence	Nima asosida hukm chiqarildi?
6.	Tavsiya	Tahrir amali va muqobil variant	edit_action, proposal, explanation	Tahrir qanday tushuntiriladi?
7.	Audit	Annotator, versiya va kelishuv	annotator_id, round, adjudication	Qaror qanday qabul qilindi?



1-rasm. Yetti sathli annotatsiya modeli manbadan auditgacha

Diagrammada sathlar ierarxiyasi ko'rsatilgan: har bir yuqori sath pastki sathning asosiy belgilariga tayanadi.

IML, USL va TER funksional sinflari

IML qatlami yozuvning me'yoriy jihatini yoritadi va grafik, orfografik hamda morfografik hodisalarga tatbiq etiladi. IML-HARF tegi harf tushishi, ortiqcha harf qo'shilishi yoki almashishini, IML-MORF esa asos bilan qo'shimcha

yoziqlashidagi nomuvofiqlikni farqlashga xizmat qiladi. Ularni kengaytiruvchi teglar sifatida IML-BOSH (bosh harf), IML-AJRAT (qo'shib/ajratib yozish), IML-CHIZIQ (chiziqcha), IML-TINISH (tinish belgisi) va IML-KOD? (grafik kodlashdagi ehtimoliy nomuvofiqlik) kiritiladi. Imloviy xato bilan grafik variantlilik ayni narsa emas: apostrof va o'zbek lotin yozuviga xos harflarning turli Unicode ko'rinishlari texnik muvofiqlik talabini tug'dirishi mumkin, biroq buni avtomatik ravishda imlo xatosi deb belgilash noto'g'ri. Shu sabab IML-KOD? yakuniy “xato” yorlig'i emas, tekshirishni talab qiluvchi ishchi flag hisoblanadi. “BERT”, DOI manzili va bibliografik rekvizitlar kabi birliklar maxsus himoyalangan segmentlar sifatida belgilanadi.

USL qatlami grammatik jihatdan to'g'ri ko'ringan, lekin ilmiy bayonning kommunikativ vazifasiga to'liq mos kelmaydigan birliklarni tahlil qiladi: USL-BAHO (dalilsiz subyektiv baho), USL-DA'VO (asossiz qat'iylashtirilgan ilmiy hukm), USL-TAKROR (mazmun qo'shmaydigan takror), USL-REFERENT (ishora birliklarining noaniq referenti), USL-MANTIQ (dalil va xulosa bog'lanishining asoslanmagani) va USL-JANR (bo'limning janriy vazifasiga nomuvofiqlik). Bu sinflar qat'iy “noto'g'ri-to'g'ri” juftlikni nazarda tutmaydi – uslubiy hukm matnning ilmiy sohasi, muallif maqsadi va kontekst bilan baholanadi. Masalan, “juda zo'r natija olindi” iborasida dalilsiz baholash xavfi bor, biroq annotator o'zi yangi ilmiy hukm (“yuqori samaradorlikka erishildi”) kiritishi mumkin emas; tahrir taklifi muallifga yo'naltirilgan izoh ko'rinishida bo'lishi lozim.

TER qatlami atamaning nafaqat grafik shaklini, balki qaysi tushunchani ifodalashini ham kuzatadi. Termin uchrasuvi uchun sirt shaklini qayd etish yetarli emas – soha kodi, concept_id, mavjud ta'rif, variantlar zanjiri va manba ishonchliligi ham qayd etiladi. Ichki sinflar: TER-ATAMA (afzal terminning muammoli tanlanishi), TER-IZCHILLIK (bir tushunchaning hujjat bo'ylab turlicha nomlanishi), TER-TA'RIF (ta'rifning yetarli ochilmasligi), TER-QISQARTMA (ochilmagan yoki chalg'ituvchi abbreviatura), TER-TARJIMA (asossiz tarjima muqobili) va TER-MA'NO (bir shaklning soha kontekstida boshqa tushunchaga ulanib ketishi).

Operatsion diagnostika va ko'p yorliqlilik muammosi

Har bir ichki teg uning asoslovchi dalili va odatiy qaror maqomi bilan bog'lanadi, bu annotatorga teglarni mexanik tanlash emas, tekshiriladigan mezon asosida tanlashga yo'naltiradi.

2-jadval. O'zbek akademik matnlari ichki teglar, asoslovchi dalillar va qaror qabul qilish mezonlari

Bosh sinf	Ichki teglar (namuna)	Asoslovchi dalil	Odatiy qaror maqomi
IML	IML-HARF, IML-MORF, IML-BOSH	Orfografik va morfografik qoida	Dalil bo'lsa: xato
IML	IML-AJRAT, IML-CHIZIQ, IML-TINISH	Qo'shib/ajratib yozish; punktuatsiya	Xato yoki tekshiruv
IML	IML-KOD?	Asl yozuv, kodlash va maket	Avval texnik tekshiruv
USL	USL-BAHO, USL-TAKROR	Ilmiy baholash va matn ixchamligi	Kontekstual tavsiya
USL	USL-DA'VO, USL-MANTIQ	Dalillash doirasi va ilmiy hukm	Muallif tasdig'i zarur
USL	USL-REFERENT, USL-JANR	Kohezionlik va bo'lim vazifasi	Kontekstual tavsiya
TER	TER-ATAMA, TER-IZCHILLIK	Tushuncha, soha va afzal nom	Dalilga bog'liq tavsiya
TER	TER-TA'RIF, TER-QISQARTMA	Ta'rif izchilligi va qisqartma	Hujjat darajasida tekshiruv
TER	TER-TARJIMA, TER-MA'NO	Soha lug'ati va tushuncha identifikatori	Noaniq holat ekspertga

Ko'p qatlamli annotatsiyada har bir hodisani bir-biridan mustaqil xato sifatida sanash noto'g'ri natija berishi mumkin. Masalan, noto'g'ri yozilgan fan termini uchun IML-HARF asosiy teg bo'lishi, shu birlikning biror concept_id ga bog'lanishi esa TER qatlamining qo'shimcha atributi sifatida saqlanishi mumkin. Demak, “bitta span – bitta xato” qat'iyligi ham, “bitta span – cheksiz ko'p mustaqil xato” erkinligi ham qabul qilinmaydi; teglarning birgalikda qo'llanishi yo'riqnomadagi asoslash shartlari bilan cheklanadi.

Har bir belgi uchun me'yoriy maqomning to'rt qiymati qo'llanadi: “tasdiqlangan xato”, “kontekstga bog'liq tavsiya”, “maqbul variant” va “ekspertga yuboriladigan noaniq holat”. Semantik xavf esa alohida past, o'rta va yuqori darajalar bilan baholanadi – raqam, foiz, inkor, modal birlik, ehtimollik bildiruvchi ifoda yoki iqtibos o'zgaradigan tahrir yuqori xavfli sifatida ko'rib chiqiladi. Bunday tasnif avtomatik ravishda tahrirning yomonligini anglatmaydi; u muallif tasdig'i zarurligini ko'rsatadi.

Annotatsiya yozuvi, audit izi va span kodlash qoidalari

Annotatsiya sxemasi faqat yorliqlar ro'yxati bo'lsa, keyinroq mustaqil ekspert uni qayta tekshira olmaydi. Har bir holat uchun quyidagi maydonlar zarur:

3-jadval. Annotatsiya maydonlari, ularning mazmuni va qabul qilinadigan qarorlar holati

Maydon	Mazmuni / qabul qilinadigan qiymat	Holati
document_id; version	Korpus manbasiga qayta yo'llanma	Majburiy
start; end; unit_scope	Matn oralig'i va tahlil ko'lam	Majburiy
primary_type; subtype	IML/USL/TER va operatsion ichki teg	Majburiy
decision_status	Xato / tavsiya / maqbul variant / noaniq	Majburiy
evidence_id; norm_scope	Manba bandi, ilmiy soha, dalil turi	Qarorga mos ravishda
concept_id; related_span	Termin tushunchasi, bog'liq uchrashuv	TER uchun shartli
edit_action; proposal	Tahrir amali yoki izohga yo'llash	Maqomga bog'liq
semantic_risk; protected_zone	Mazmunni o'zgartirish xavfi va himoya	Majburiy tekshiruv
annotator_id; scheme_version	Kim va qaysi yo'riqnoma bilan belgilagan	Majburiy
adjudication; audit_log	Kelishuv, qaror va tarix	Tajribadan so'ng

Annotatsiyalashning barcha versiyalari kuzatilishi shart: dastlabki ikki annotator yozuvi o'zgaras xom qatlamda, ekspert qarori esa alohida adjudicated qatlamda saqlanadi. Belgilash chegarasi nomuvofiqlikni diagnostika qilish uchun zarur eng kichik uzluksiz matn oralig'ini qamrashi lozim. Tahrir amallari beshta: insertion (qo'shish, nol uzunlikdagi pozitsiya sifatida), delete (olib tashlash), replace (almashtirish), split va merge (qayta segmentatsiya). Biroq “tahrir amali” va “lingvistik xato turi” turli o'lchamlardir – bitta replace morfografik, leksik yoki terminologik sababga ko'ra yuz berishi mumkin.

4-jadval. Annotatsiyalash jarayonida tahrir amallari, POS teglari va terminologik birliklar bo'yicha annotator qarorlari hamda ehtiyot choralari

Namuna birlik	Teg / hukm	Annotator qarori va ehtiyot sharti
“algoritim”	IML-HARF / xato	“algoritim”; ilmiy terminning asl shakli tekshiriladi
“ma'lumotlar”	IML-MORF / xato	“ma'lumotlar”; OCR yoki satr nuqsoni rad etiladi
“Ushbu ishda mazkur ishda”	USL-TAKROR / tavsiya	Ortiqcha takrorni ajratish; mazmunni saqlash
“Juda zo'r natija”	USL-BAHO / tavsiya	Dalil va o'lchovga oid izoh so'rash
“Bu natija”	USL-REFERENT / tekshiruv	Oldingi gaplarda aniq referent bor-yo'qligini ko'rish
“Barcha holatlarda isbotlandi”	USL-DA'VO / yuqori xavf	Sinov qamrovini tekshirish; avtomatik yumshatmaslik
“NLP” birinchi uchrashuvi	TER-QISQARTMA / shartli	Oldin ochilgan-ochilmaganini hujjatda tekshirish
“masofaviy” ↔ “onlayn”	TER-MA'NO / noaniq	Avval tushuncha farqini aniqlash; sinonim demaslik
DOI, formulalar, kod	HIMOYA / tahrirga yopiq	Normativ dalilsiz o'zgartirmaslik

Izoh: jadvaldagi barcha birliklar uslubiy-tahliliy misollar sifatida tuzilgan.

Yo'riqnoma, qaror daraxti va bahsli holatlarni hal qilish

Yo'riqnoma annotatorning shaxsiy didini me'yoriy hukmga almashtirib yubormaslik uchun ishlatiladi. Unda har bir tegning ta'rifi, qo'llash sharti, qo'llamaslik sharti, tipik misollar, qarshi misollar, span chegarasi, dalilning ustuvorlik tartibi va bahsli holatni ekspertga yuborish yo'li belgilanadi. Ikki malakali annotator bir xil matnni mustaqil ko'rib, bir xil qarorga kelishi uchun yo'riqnomaning operatsion savollari yetarli darajada aniqlashtiriladi.

Birlamchi qaror daraxti quyidagi olti bosqichda tasavvur qilinadi:

1. Birlik asl muallif matnidan olinganmi yoki ajratib olishda yuzaga kelgan texnik buzilishmi, aniqlanadi.
2. Birlik himoyalangan hududga kiradimi, tekshiriladi.
3. Nomuvofiqlik mavjudligi uchun normativ yoki diskursiv dalil bor-yo'qligi baholanadi.
4. Uning IML/USL/TER sinfi hamda ichki tegi tanlanadi.
5. Majburiy tuzatish, ixtiyoriy tavsiya, maqbul variant va noaniq holat bir-biridan farqlanadi.
6. Semantik xavf, ehtimoliy tahrir yoki izoh va manba ko'rsatkichi kiritiladi.



2-rasm. Qaror daraxtining olti qadami

Ushbu ketma-ketlik annotatsiya uchun taklif etilgan metodik qoida bo'lib, hozircha validatsiya qilingan qaror daraxti sifatida talqin etilmaydi.

Taklif etilayotgan adjudikatsiya ikki bosqichli: birinchi bosqichda annotatorlar kelishmovchilikni o'z dalillari bilan hujjatlashtiradi, ikkinchisida mustaqil ekspert yoki metodik moderator ko'rsatma asosida yakuniy holatni tasdiqlaydi. Bahs yechilmasa, yozuv “noaniq” maqomida qoldiriladi va uni majburiy tuzatish uchun o'quv namunasiga kiritmaslik ko'zda tutiladi.

Oltin standart annotatsiya – mutlaq haqiqiylik emas, balki belgilangan yo'riqnoma hamda ekspert dalillariga ko'ra kelishilgan referens qarorlar majmuasidir. Uni yaratishdan avval mustaqil annotatorlar yozuvi hisoblanadi, so'ng yechilgan kelishmovchiliklardan adjudicated gold ajratiladi; gold yaratilganidan keyin annotatorlar o'rtasidagi oldingi mustaqil tafovutlar yo'qotilmasligi kerak. Shu bois raw_A, raw_B va gold versiyalari alohida saqlanadi.

Validator texnik yaxlitlikni ta'minlaydi: har bir span chegarasi matn uzunligi doirasida ekanini, bo'sh bo'lmagan interval original_fragment bilan mosligini, kategoriya yorlig'i ruxsat etilgan sxemaga tegishli ekanini va manba versiyasi o'zgarmaganini nazorat qiladi. Lingvistik mazmunni ekspert, texnik yaxlitlikni esa deterministik validator baholaydi.

Xulosa. Taklif qilingan annotatsiyalash sxemasi til tasnifini korpusning empirik resurslariga o'tkazishning operatsion mexanizmini yaratadi. Uning nazariy yangiligi uchta bosh sinfga yangi nom qo'yishda emas, xato chegarasi, tahrir maqomi, fan sohasi, dalil manbasi va muallif ilmiy hukmini saqlash shartlarini bitta tekshiriladigan annotatsiya yozuvida bog'lashdadir. Shu asosda qat'iy imloviy me'yor buzilishi bilan ilmiy uslub bo'yicha tavsiya yoki terminologik variant bir xil turdagi “to'g'rilash” misoliga aylantirilmaydi. Bu farqlash keyingi avtomatik modelni nafaqat tuzatish kiritishga, balki qachon tavsiya bermaslik zarurligini ham o'rganishga yo'naltiradi.

Demak, ishonchli annotatsiya oddiy texnik tayyorlov emas, balki ilmiy baholanadigan lingvistik referens qatlam hisoblanadi. Sxemaning amaliy ishonchliligi – annotatorlararo kelishuv ko'rsatkichlari, ularning talqin chegaralari va validatsiya protokoli – alohida metodologik masala bo'lib, keyingi tadqiqotda batafsil ko'rib chiqiladi.



Foydalanilgan adabiyotlar ro'yxati:

1. Pustejovsky J., Stubbs A. Natural Language Annotation for Machine Learning. Sebastopol: O'Reilly Media; 2012.
2. International Organization for Standardization. ISO 24612:2012. Language Resource Management – Linguistic Annotation Framework (LAF). Geneva: ISO; 2012.
3. Bryant C., Felice M., Briscoe T. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. Proceedings of ACL. 2017:793–805. doi:10.18653/v1/P17-1074.
4. O'zbekiston Respublikasi Vazirlar Mahkamasi. O'zbek tilining asosiy imlo qoidalarini tasdiqlash haqida. 1995-yil 24-avgust, 339-son qaror.
5. International Organization for Standardization. ISO 704:2022. Terminology Work – Principles and Methods. Geneva: ISO; 2022.
6. Unicode Consortium. Unicode Standard Annex #29: Unicode Text Segmentation. Revision 49. 2026.