

Oriental Journal of Philology**ORIENTAL JOURNAL OF PHILOLOGY**

journal homepage:

<http://www.supportscience.uz/index.php/ojp/about>**A MULTI-LAYER ANNOTATION MODEL FOR AN ENGLISH–UZBEK CORPUS OF COLLOQUIAL UNITS****Istorahon Sotvoldiyeva***PhD in Philological Sciences**Lecturer at Andijan State Institute of Foreign Languages**E-mail: istorasotvoldiyeva@gmail.com**Science ID: FAN-0925-0049**Andijan, Uzbekistan***ABOUT ARTICLE**

Key words: corpus linguistics, English–Uzbek corpus, colloquial unit, annotation, pragmatics, functional equivalent, NLP, artificial intelligence.

Received: 18.09.26**Accepted:** 19.09.26**Published:** 20.09.26

Abstract: The article proposes a multi-layer annotation model for an English–Uzbek corpus of colloquial units. Existing Uzbek national, parallel, morphological and syntactic resources and international multiword-expression corpora are reviewed to identify a gap in pragmatics-aware bilingual description. The proposed record combines an English expression, a short authentic context and an Uzbek functional equivalent with layers for form, source metadata, contextual meaning, pragmatic function, register, evaluative stance, participant relations, AI suggestions and expert decisions. A quality-control workflow based on independent checking, adjudication and licensing constraints is also outlined.

INGLIZ–O‘ZBEK KOLLOKVIAL KORPUSINI YARATISHDA KO‘P QATLAMLI ANNOTATSIYA MODEL**Istorahon Sotvoldiyeva***Andijon davlat chet tillari instituti f.f.d. (PhD), o‘qituvchisi**E-mail: istorasotvoldiyeva@gmail.com**Ilmiy identifikatsiya raqami: FAN-0925-0049**Andijon, O‘zbekiston***MAQOLA HAQIDA**

Kalit so‘zlar: korpus lingvistikasi, ingliz–o‘zbek korpusi, kollokvial birlik, annotatsiya, pragmatik belgi, funksional muqobil, NLP, sun‘iy intellekt.

Annotatsiya: Maqolada ingliz–o‘zbek kollokvial korpusini yaratish uchun semantik, uslubiy va pragmatik axborotni birlashtiruvchi ko‘p qatlamli annotatsiya modeli ishlab chiqiladi. O‘zbek tili uchun mavjud milliy,

parallel, morfologik va sintaktik resurslar hamda xalqaro ko'p so'zli birliklar korpuslari tahlil qilinib, kollokvial birliklarni funksional muqobil darajasida tavsiflashga oid bo'shliq ko'rsatiladi. Korpus birligi sifatida inglizcha ifoda, qisqa autentik kontekst va o'zbekcha funksional muqobilning yaxlit yozuvi taklif etiladi. Model formal pasport, manba metama'lumoti, kontekstdagi ma'no, pragmatik vazifa, registr, emotsional-baholovchi bo'yoq, ishtirokchilar munosabati, o'zbekcha funksional muqobil, AI taklifi va ekspert qarori kabi qatlamlarni qamrab oladi. Shuningdek, annotatorlararo muvofiqlik, ekspert tekshiruv va litsenziyaviy cheklovlarni hisobga oluvchi sifat nazorati sxemasi tavsiya etiladi.

МНОГОУРОВНЕВАЯ МОДЕЛЬ АННОТАЦИИ ПРИ СОЗДАНИИ АНГЛО-УЗБЕКСКОГО КОРПУСА РАЗГОВОРНЫХ ЕДИНИЦ

Исторахон Сотволдиева

PhD по филологическим наукам, преподаватель Андижанского государственного института иностранных языков

E-mail: istorasotvoldiyeva@gmail.com

Научный идентификационный номер: FAN-0925-0049

Андижан, Узбекистан

О СТАТЬЕ

Ключевые слова: корпусная лингвистика, англо-узбекский корпус, разговорная единица, аннотация, прагматика, функциональный эквивалент, NLP, искусственный интеллект.	Аннотация: В статье предлагается многоуровневая модель аннотации для англо-узбекского корпуса разговорных единиц. На основе анализа существующих узбекских морфологических, синтаксических и параллельных ресурсов, а также международных корпусов многословных выражений определяется методологический пробел в прагматически ориентированном двуязычном описании. Модель объединяет форму единицы, источник и контекст, значение, прагматическую функцию, регистр, оценочность, отношения между участниками, функциональный эквивалент на узбекском языке и документирование решений ИИ и эксперта.
--	--

Kirish. Korpus lingvistikasi til birliklarini alohida misollar emas, tekshiriladigan manba va kontekst bilan tizimli ravishda tavsiflash imkonini beradi. O'zbek tili uchun so'nggi yillarda raqamli resurslar sezilarli kengaydi: Milliy korpusda parallel qidiruv bo'limi mavjud, ingliz–

o'zbek parallel korpusini lingvistik va ekstralingvistik teglash tamoyillari bo'yicha metodik ishlar olib borilgan [1; 97–109-b.]. Universal Dependencies doirasida o'zbek tili uchun qo'lda tekshirilgan treebanklar yaratildi; 2025-yilda e'lon qilingan Uzbek-UT 500 gap va 5850 tokenni, 2026-yildagi UzUDT esa 684 gap va 7582 tokenni qamrab oladi [3; 1–6-b.; 4; 11642–11649-b.]. UzbekPOS resursida qariyb 4,5 ming gap va 53 mingdan ortiq token/teg juftligi professional annotatorlar tomonidan belgilangan [5]. Bu ishlar o'zbek tilining hisoblash resurslari tez rivojlanayotganini ko'rsatadi.

Shu bilan birga, mavjud resurslarning asosiy kuchi morfologiya, so'z turkumi, sintaktik bog'lanish, umumiy parallel moslik yoki matn qidiruviga qaratilgan. Kundalik muloqotdagi kollokvial birlik uchun "kinga, qanday vaziyatda, qanday ohangda va nima maqsadda aytilgan?" savollarini muntazam belgilaydigan ingliz–o'zbek maxsus qatlam alohida ehtiyoj sifatida qolmoqda. Avvalgi qiyosiy tadqiqotda inglizcha kollokvial birliklarning o'zbekcha mosliklari semantik jihatdan ko'rib chiqilgan [9; 5–9-b.], ammo korpus yozuvi darajasida pragmatik funksiya, registr, interaksion munosabat va AI/ekspert qarorini birlashtiruvchi sxema ishlab chiqilmagan.

Maqolaning maqsadi ingliz–o'zbek kollokvial korpusini qurishda qo'llash mumkin bo'lgan ko'p qatlamli annotatsiya modelini ishlab chiqishdir. Vazifalar quyidagilardan iborat: mavjud o'zbek va xalqaro korpus resurslarining annotatsiya yo'nalishini qiyoslash; kollokvial birlik uchun minimal va kengaytirilgan yozuv tarkibini belgilash; semantik, pragmatik va sotsiolingvistik qatlamlarni bir sxemada birlashtirish; ekspert tekshiruvi va inson nazoratidagi AI jarayoni uchun sifat nazorati tamoyillarini taklif etish.

Mavzuga oid adabiyotlar tahlili. Ko'p so'zli birliklar NLP uchun murakkab obyekt hisoblanadi, chunki ularning ma'nosi komponentlar yig'indisiga to'liq teng bo'lmasligi, struktur jihatdan o'zgaruvchan va kontekstga bog'liq bo'lishi mumkin. Constant va hammualliflar MWElarni aniqlash va qayta ishlashni parsing hamda mashina tarjima bilan bog'liq markaziy vazifa sifatida ko'rsatib, maxsus annotatsiyalangan resurslarning ahamiyatini asoslaydi [6; 837–892-b.]. Dialog ma'lumotlari bo'yicha DiaBLa ingliz–fransuz spontan yozma muloqotida gap darajasidagi tarjima bilan bir qatorda butun dialog kontekstini saqlash kerakligini ko'rsatadi [7; 635–660-b.].

2026-yilda Han va hammualliflar taqdim etgan AlphaMWE korpusi parallel MWE resurslarini yaratishda MT-assisted va human-in-the-loop yondashuvini birlashtiradi. Unda avtomatik tarjima inson post-editing, qo'lda MWE annotatsiyasi va ikkinchi sifat tekshiruvi bilan to'ldirilgan; mualliflar yuqori sifatli bilingval MWE ma'lumotlari hali kamligini alohida ta'kidlaydi [8]. Ushbu tajriba kollokvial korpus uchun ham muhim: AI katta hajmdagi

nomzodlarni tez topishi mumkin, biroq kontekst, registr va pragmatik ma'noni yakuniy belgilash ekspert tekshiruvini talab qiladi.

O'zbek tilidagi mavjud resurslar esa boshqa zarur qatlamlarni ko'rsatadi. Elov va Amirqulov parallel korpusda material tanlash, alignment, POS tagging va XML kodlash kabi lingvistik hamda texnik tamoyillarni muhokama qiladi [1]. Universal Dependencies resurslari aniq annotatsiya yo'riqnomasi va ko'p annotatorli tekshiruvning qayta qo'llanadigan modelini namoyish etadi [3; 4]. UzbekPOSda har bir annotatsiya kamida ikki annotator tomonidan tekshirilgani sifat nazoratining muhimligini ko'rsatadi [5]. Biroq morfosintaktik to'g'rilik kollokvial tarjimaning pragmatik mosligini avtomatik ta'minlamaydi. Shu sababli yangi resurs yuqoridagi tajribalarni pragmatik va tarjimaviy qatlamlar bilan boyitishi lozim.

Tadqiqot metodologiyasi. Maqola resurs dizayni va annotatsiya metodologiyasiga asoslangan nazariy-amaliy tadqiqotdir. Tahlil birligi sifatida alohida leksema emas, balki "inglizcha kollokvial birlik + qisqa autentik kontekst + o'zbekcha funksional muqobil" uchligi qabul qilindi. Birlik suhbat formulasi, diskurs markeri, norasmiy ko'p so'zli ifoda, kundalik idiomatik birikma yoki yengil sleng bo'lishi mumkin. Tanlashda uch mezon ustuvor: real muloqotdagi funksionallik, zamonaviy qo'llanishning tekshirilishi va o'zbekchallashtirishda kontekstga sezgir qaror talab qilishi.

Resurs yaratishda ochiq yoki ilmiy foydalanishga ruxsat etilgan og'zaki manbalar, podkast va intervyu transkriptlari, ikki yoki undan ortiq ishtirokchi qatnashgan tabiiy muloqotlar afzal ko'riladi. To'liq transkriptni ko'chirish o'rniga birlikni talqin qilish uchun zarur qisqa kontekst, manba turi, sana va identifikatsiya ma'lumoti saqlanadi. Bu usul lingvistik tekshiriluvchanlik bilan mualliflik huquqi ehtiyotkorligini muvozanatlashtiradi. Har bir yozuvda AI nomzod yoki dastlabki tarjima taklif qilishi mumkin, ammo yakuniy qabul, tahrir yoki rad qarori tadqiqotchi/ekspertga tegishli bo'ladi.

Tahlil va natijalar. Birinchi natija mavjud resurslarning bir-birini to'ldirishi, lekin kollokvial-pragmatik maqsad uchun tayyor holda yetarli emasligini ko'rsatdi. Parallel korpuslar til juftligi bo'yicha alignment beradi; UD treebanklari morfosintaktik strukturani; UzbekPOS so'z turkumi va morfologik teglashni; dialog va MWE korpuslari esa interaksion kontekst hamda ko'p so'zli birlikka xos qiyinchiliklarni ochadi. Yangi ingliz-o'zbek kollokvial korpusi ushbu imkoniyatlarni funksional tarjima va pragmatik belgilash atrofida birlashtirishi kerak.

1-jadval. Mavjud resurslar va kollokvial korpus ehtiyoji

Resurs	Asosiy turi	Kuchli tomoni	Kollokvial tadqiqot uchun cheklov
Uzbek Milliy korpusi	milliy/parallel qidiruv	lemma, konkordans, parallel matn	maxsus kollokvial-pragmatik yozuv markaziy emas

Uzbek-UT / UzUDT	UD treebank	POS, morfologiya, dependency	tarjima va pragmatik muqobilni bermaydi
UzbekPOS	multi-domain POS dataset	53 mingdan ortiq token/teg juftligi	interaksion funksiya annotatsiyasi yo‘q
DiaBLa	bilingual dialogue	spontan dialog, kontekst, MT bahosi	ingliz–fransuz juftligi
AlphaMWE	multilingual parallel MWE	MWE alignment, MT + human checking	kollokvial pragmatika alohida markaz emas
Taklif etilayotgan model	ingliz–o‘zbek kollokvial	semantika + pragmatika + registr + muqobil	maxsus kontekstga sezgir ikki tilli qatlam

Ikkinchi natija korpus yozuvini faqat “English — Uzbek” ikki ustunli jadval ko‘rinishida qurish yetarli emasligini ko‘rsatdi. Kollokvial birlikni qayta ishlash uchun kamida o‘nta axborot qatlamini ajratish maqsadga muvofiq. Formal qatlam lema, variant va struktur tipni; manba qatlamida platforma, sana va kontekst identifikatori; semantik qatlam kontekstdagi ma‘noni; pragmatik qatlam nutq funksiyasini; registr qatlam uslubiy darajani; evaluativ qatlam ijobiy/salbiy/neytral bo‘yoqni; relatsion qatlam suhbatdoshlar o‘rtasidagi masofani; tarjima qatlam o‘zbekcha funksional muqobil va alternativlarni; AI izi dastlabki avtomatik taklifni; ekspert qatlami esa yakuniy qaror, tahrir va ishonch darajasini saqlaydi.

2-jadval. Ingliz–o‘zbek kollokvial yozuvining ko‘p qatlamli annotatsiya modeli

Qatlam	Saqlanadigan ma’lumot	Asosiy vazifa
1. Formal pasport	lema, variant, so‘z turkumi/struktur tip	birlik shaklini normallashtirish
2. Manba va kontekst	manba, sana, qisqa suhbat parchasi	tekshiriluvchanlik
3. Kontekstdagi ma‘no	qisqa semantik izoh	polisemiyani ajratish
4. Pragmatik vazifa	rozilik, rad, yumshatish, kinoya va h.k.	kommunikativ maqsadni belgilash
5. Registr	neytral, norasmiy, juda norasmiy	uslubiy moslik
6. Evaluativ bo‘yoq	ijobiy, salbiy, neytral/aralash	munosabatni ko‘rsatish
7. Ishtirokchilar munosabati	yaqinlik, maqom, adresat turi	sotsiolingvistik kontekst
8. O‘zbekcha funksional muqobil	asosiy va alternativ variantlar	tabiiy tarjima
9. AI izi	AI taklifi va izohi	avtomatik yordamni audit qilish
10. Ekspert qarori	qabul/tahrir/rad, ishonch darajasi	sifat va javobgarlik

Uchinchi natija pragmatik belgilash uchun qat’iy annotatsiya yo‘riqnomasi zarurligidir. “Kinoya”, “norozilik”, “yumshatish” kabi yorliqlar intuitiv qo‘yilsa, annotatorlararo tafovut oshadi. Shu sababli har bir kategoriya uchun ta’rif, ijobiy va salbiy mezon, minimal juft misollar hamda bahsli holatlarda qaror daraxti berilishi kerak. Pilot bosqichda bir xil kichik to‘plamni ikki yoki uch annotator mustaqil belgilashi, keyin tafovutlar muhokama va adjudikatsiya orqali hal qilinishi maqsadga muvofiq. Korpusning ma’lum qismi doimiy ravishda ikkilamchi tekshiruvdan

o'tkazilishi resursni kengaytirishda izchillikni saqlaydi. Bu tamoyil UzUDT va UzbekPOSdagi ko'p annotatorli nazorat tajribasi bilan hamohang [4; 5].

To'rtinchi natija AI yordamidan foydalanishda "natija" emas, "qaror izi"ni saqlash muhimligini ko'rsatdi. Agar tizim kollokvial birlikni nomzod sifatida ajratsa yoki o'zbekcha tarjima taklif qilsa, korpusda faqat yakuniy variantni qoldirish tadqiqot uchun qimmatli ma'lumotni yo'qotadi. AI taklifi, inson tahriri, rad sababi va yakuniy variant yonma-yon saqlansa, keyinchalik qaysi birliklar avtomatik tizimlar uchun qiyin ekanini, qaysi xato turlari takrorlanishini va inson ekspertizasi qayerda hal qiluvchi bo'lganini empirik o'rganish mumkin. AlphaMWE tajribasida ham MT natijasini professional post-editing va qayta tekshirish bilan birlashtirish resurs sifatining tayanchi sifatida ishlatilgan [8].

Modelning amaliy chiqishi elektron bazada bir nechta qidiruv o'qini yaratadi: foydalanuvchi birlik shakli, pragmatik vazifa, registr, evaluativ bo'yoq yoki o'zbekcha muqobil bo'yicha filtrlashi mumkin. Masalan, "rad etish + norasmiy" filtri no way, I'm good, not really kabi turli shakllarning bir kommunikativ maydonda qanday ifodalanishini ko'rsatadi. Bu yondashuv korpusni oddiy lug'at ro'yxatidan farqlab, tadqiqot, tarjima va ta'lim uchun qayta ishlatiladigan annotatsiyalangan resursga aylantiradi.

Muhokama. Ko'p qatlamli annotatsiya modeli resursning ilmiy qiymatini oshiradi, biroq annotatsiya hajmi bilan amaliy bajarilish o'rtasida muvozanat saqlanishi kerak. Har bir birlikka haddan tashqari ko'p maydon biriktirish kichik jamoa uchun ish tezligini keskin pasaytirishi mumkin. Shu sababli korpusda "majburiy" va "kengaytirilgan" maydonlarni ajratish maqsadga muvofiq. Lema, kontekst, manba, kontekstdagi ma'no, pragmatik vazifa, registr va o'zbekcha funksional muqobil majburiy yadro bo'lsa, emotsional bo'yoq, ishtirokchilar munosabati, AI izi va ishonch darajasi kengaytirilgan qatlam sifatida bosqichma-bosqich to'ldirilishi mumkin. Bu yechim pilot bosqichda tezlikni, keyingi bosqichda esa ilmiy chuqurlikni saqlaydi.

Modelning yana bir muhim tomoni interoperabellikdir. Agar maydon nomlari, teglar va eksport formati oldindan standartlashtirilsa, korpus keyinchalik boshqa NLP vositalari, konkordans dasturlari yoki ta'lim platformalariga ulanishi osonlashadi. Shu bilan birga, korpusga kiruvchi og'zaki materiallar platforma va janr bo'yicha muvozanatli tanlanmasa, bir turdagi podkast yoki yosh guruhining nutqi butun zamonaviy Amerika ingliz tilining vakili sifatida noto'g'ri umumlashtirilishi mumkin. Demak, annotatsiya sifati bilan bir qatorda tanlovning reprezentativligi, manba huquqi va versiyalash tamoyillari ham korpus dizaynining ajralmas qismi bo'lishi kerak.

Xulosa. O'zbek tili uchun yaratilayotgan zamonaviy treebank, POS va parallel korpuslar muhim infratuzilma yaratmoqda, biroq ingliz-o'zbek kollokvial birliklarni kontekst, pragmatik vazifa, registr va funksional muqobil bilan birga tavsiflash alohida resurs dizaynini talab qiladi.

Taklif etilgan oʻn qatlamli model formal pasportdan ekspert qarorigacha boʻlgan barcha muhim maʼlumotni bir yozuvda birlashtiradi. Uning ilmiy afzalligi birlikning shakli va tarjimasini emas, tarjima qarorini yuzaga keltirgan kontekstual-pragmatik omillarni ham saqlashidadir. Keyingi bosqichda annotatsiya yoʻriqnomasi pilot toʻplamda sinovdan oʻtkazilib, annotatorlararo muvofiqlik, maʼlumot kiritish vaqti hamda foydalanuvchi qidiruv ssenariylari boʻyicha baholanishi lozim.

Foydalanilgan adabiyotlar roʻyxati:

1. Elov B., Amirqulov M. Steps of creating a tagged parallel corpus of the Uzbek-English languages // Uzbekistan Language and Culture. – 2024. – Vol. 5, No. 1. – P. 97–109.
2. Uzbek National Corpus. Parallel corpus and corpus search resources. – URL: <https://uzbekcorpus.uz/> (murojaat sanasi: 22.09.2026).
3. Akhundjanova A., Talamo L. Universal Dependencies Treebank for Uzbek // Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025). – Tallinn, 2025. – P. 1–6.
4. Matlatipov S.G., Aripov M. UzUDT: Uzbek Universal Dependencies Treebank // Proceedings of the Fifteenth Language Resources and Evaluation Conference. – Palma de Mallorca: ELRA, 2026. – P. 11642–11649. – DOI: 10.63317/2uedjqezjxn5.
5. Sharipov M., Kuriyozov E., Vičič J. UzbekPOS: A multi-domain dataset for Uzbek part-of-speech tagging // Data in Brief. – 2026. – Vol. 66. – Article 112640. – DOI: 10.1016/j.dib.2026.112640.
6. Constant M., Eryiğit G., Monti J., van der Plas L., Ramisch C., Rosner M., Todirascu A. Multiword Expression Processing: A Survey // Computational Linguistics. – 2017. – Vol. 43, No. 4. – P. 837–892. – DOI: 10.1162/COLI_a_00302.
7. Bawden R., Bilinski E., Lavergne T., Rosset S. DiaBLa: a corpus of bilingual spontaneous written dialogues for machine translation // Language Resources and Evaluation. – 2021. – Vol. 55. – P. 635–660. – DOI: 10.1007/s10579-020-09514-4.
8. Han L., Mohamed N.H., Rassem M., Jones G.J.F., Smeaton A.F., Nenadic G. Towards a resource for multilingual lexicons: an MT assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation // Language Resources and Evaluation. – 2026. – Vol. 60. – Article 33. – DOI: 10.1007/s10579-025-09876-7.
9. Sotvoldieva I.L. Contrastive analysis of cross-language colloquial correspondence in English and Uzbek Languages // Texas Journal of Philology, Culture and History. – 2024. – Vol. 33. – P. 5–9. – DOI: 10.62480/tjpch.2024.vol33.pp5-9.
10. Baker M., Jones H. In Other Words: A Coursebook on Translation. 4th ed. – London: Routledge, 2026. – 410 p. – DOI: 10.4324/9781003440857.