“ULUG‘BEK XAZINASI” ROMANINI RAQAMLASHTIRISH, SEGMENTATSIYALASH HAMDA UNING INGLIZCHA PARALLEL BAZASINI YARATISH

https://doi.org/10.70728/phil.v05.i08.031

Gulmira Komilova
ORCID: 0009-0009-6585-975X
komilovaguli0207@gmail.com
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
Toshkent xalqaro moliyaiy boshqaruv va texnologiyalari universiteti

Annotatsiya: Maqolada Odil Yoqubovning “Ulug‘bek xazinasi” romanini raqamli gumanitar fanlar va korpus lingvistikasi tamoyillari asosida raqamlashtirish, strukturaviy segmentatsiyalash hamda uning inglizcha tarjimasi bilan o‘zbek-ingliz parallel bazasini yaratishning formal modeli ishlab chiqiladi. Tadqiqotning asosiy muammosi badiiy-tarixiy matnda gap va abzas chegaralarining tarjimada muntazam saqlanmasligi, dialoglar, muallif nutqi, tarixiy realiyalar, atoqli otlar, arxaik birliklar va izohlovchi tarjima tufayli 1:1 moslikning tez-tez buzilishidir. Shu sabab ikki bosqichli yondashuv taklif etiladi: avval matnlar nashr-qism-bob-abzas-gap-token ierarxiyasida segmentlanadi, so‘ng segmentlar 1:1, 1:n, n:1, n:m, 1:0 va 0:1 turlarida moslashtiriladi. Maqolada barqaror segment identifikatorlari, xom va normallashtirilgan matn qatlamlari, TEI P5 bilan mos stand-off belgilash, relyatsion ma’lumotlar modeli, versiyalash va provenance nazorati, OCR xatolarini baholash, segmentatsiya F1 ko‘rsatkichi, alignment precision/recall/F1 va annotatorlararo kelishuv mezonlari tavsiya etiladi. Natijada “Ulug‘bek xazinasi” asosida keyingi morfologik, leksik-semantik, nomlangan obyektlar, tarixiy realiyalar hamda tarjima transformatsiyalarini ko‘p qatlamli annotatsiyalashga tayyor, qayta tekshiriladigan va kengaytiriladigan o‘zbek-ingliz parallel korpus protokoli shakllantiriladi.

Kalit so‘zlar: parallel korpus, korpus lingvistikasi, raqamlashtirish, OCR, segmentatsiya, sentence alignment, TEI P5, badiiy tarjima, tarixiy realiya, “Ulug‘bek xazinasi”, ma’lumotlar bazasi

DIGITIZING AND SEGMENTING THE NOVEL “THE TREASURE OF ULUGBEK” AND BUILDING ITS ENGLISH PARALLEL DATABASE

Gulmira Komilova
ORCID: 0009-0009-6585-975X
komilovaguli0207@gmail.com

(cc) BY

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi
Tashkent International University of Financial Management and Technologies

Abstract: This article proposes a formal corpus-linguistic workflow for digitizing and structurally segmenting Odil Yakubov's historical novel "Ulug'bek xazinasi" and for building an Uzbek-English parallel database with its English translation. The central methodological problem is that literary translation does not preserve paragraph and sentence boundaries consistently: dialogue, narrator speech, culture-specific items, proper names, archaisms, explicitation and compression frequently break one-to-one correspondence. The proposed workflow therefore separates segmentation from alignment. Both language versions are first represented through a document-part-chapter-paragraph-sentence-token hierarchy with stable identifiers; the resulting units are then aligned through 1:1, 1:n, n:1, n:m, 1:0 and 0:1 relations. The study specifies raw and normalized text layers, TEI P5-compatible stand-off linking, a relational data model, versioning and provenance controls, and quality-assurance metrics for OCR, boundary detection, alignment and inter-annotator agreement. The resulting protocol is designed as a reproducible foundation for subsequent multilayer Abstract of morphology, named entities, historical realia, lexical-semantic relations and translation transformations.

Keywords: parallel corpus, corpus linguistics, digitization, OCR, segmentation, sentence alignment, TEI P5, literary translation, historical realia, database

1. Kirish

Raqamli gumanitar fanlar, korpus lingvistikasi va tarjima texnologiyalarining kesishgan nuqtasida yaratiladigan parallel korpuslar badiiy matnni faqat elektron shaklda saqlash vositasi emas, balki til birliklari, tarjima strategiyalari, semantik munosabatlar va madaniy xususiyatlarni tizimli kuzatish imkonini beradigan tadqiqot infratuzilmasidir. Ayniqsa, o'zbek tilining ingliz tili bilan parallel resurslari cheklangan sharoitda milliy adabiyot namunalarini standartlashtirilgan formatda raqamlashtirish va moslashtirish kompyuter lingvistikasi, tarjimashunoslik, leksikografiya hamda sun'iy intellekt uchun bir xil darajada ahamiyatlidir [3; 4; 8].

Odil Yoqubovning "Ulug'bek xazinasi" romani bunday resurs uchun murakkab, lekin ilmiy jihatdan sermahsul materialdir. Asarda tarixiy shaxslar, saroy va ilmiy muhitga oid terminlar, arxaik va kitobiy qatlam, turkiy, forsiy-arabiy birliklar, dialog va ichki monolog, muallif nutqi hamda tarixiy-madaniy realiyalar zich qo'llanadi. Adabiyotshunoslikda roman tarixiy xotira, obraz, xronotop va milliy kolorit nuqtayi nazaridan o'rganilgan; tarjimashunoslikda esa ayrim realiyalar va lingvistik transformatsiyalar tahlil qilingan [1; 2; 16; 17]. Biroq asarning o'zbekcha va inglizcha

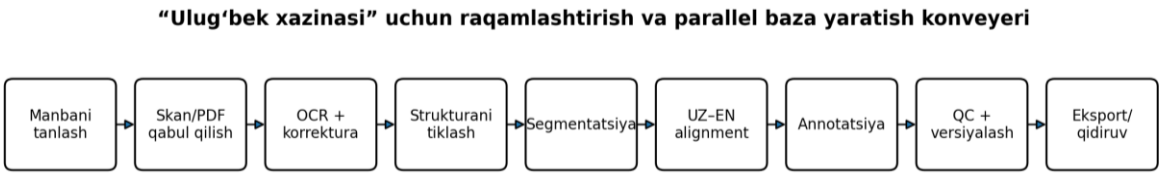
(cc) BY

versiyasini bir xil identifikatorlar, tekshiriladigan alignment aloqalari va ko‘p qatlamli annotatsiyaga tayyor ma’lumotlar modeli asosida birlashtiradigan korpus protokoli alohida ishlab chiqishni talab qiladi.

2026-yilda “The Treasure of Ulugbek” nomi bilan inglizcha nashrlar qayd etilgani roman asosida yangi o‘zbek-ingliz parallel resurs yaratish imkoniyatini kuchaytirdi [2; 18]. Shu bilan birga, 2026-yilgi bibliografik qaydlarda nashr varianti, tarjimonlar va nashriyot ma’lumotlari bir xil ko‘rinishda berilmaydi. Korpus yaratishda bu holat oddiy bibliografik tafsilot emas: matnning qaysi versiyasi asos qilib olingani keyingi segmentlar, sahifa mosliklari, tuzatishlar va tarjima tahlilining takrorlanuvchanligini bevosita belgilaydi. Shuning uchun mazkur tadqiqotda nashr identifikatsiyasi va provenance nazorati ma’lumotlar modelining ajralmas qismi sifatida qaraladi.

Tadqiqotning maqsadi - “Ulug‘bek xazinasi” romanini raqamlashtirish, strukturaviy va lingvistik segmentatsiyalash hamda inglizcha tarjima bilan parallel bazaga birlashtirish uchun takrorlanuvchan, standartlarga mos va ko‘p qatlamli annotatsiyaga tayyor modelni ishlab chiqish. Maqsadga erishish uchun quyidagi vazifalar belgilandi: (1) nashr va fayl darajasidagi manba nazoratini formalizatsiyalash; (2) OCR va korrektura jarayonini xom hamda normallashtirilgan qatlamlarga ajratish; (3) qism, bob, abzas, gap, dialog bloki va token uchun barqaror segment identifikatorlarini belgilash; (4) o‘zbekcha va inglizcha segmentlar o‘rtasidagi ko‘p turli alignment munosabatlarini modellashtirish; (5) TEI P5 va relyatsion ma’lumotlar bazasi o‘rtasida mos eksport/import mexanizmini belgilash; (6) sifat nazorati uchun o‘lchanadigan mezonlarni ishlab chiqish.

Tadqiqotning ilmiy yangiligi shundan iboratki, badiiy-tarixiy matn uchun segmentatsiya va alignment bir jarayon sifatida emas, balki mustaqil, lekin bog‘langan ikki qatlam sifatida formalizatsiya qilinadi; segment identifikatori matnning pozitsiyasidan tashqari nashr versiyasi bilan bog‘lanadi; 1:1 dan tashqari 1:n, n:1, n:m, 1:0 va 0:1 mosliklar birlamchi obyekt sifatida saqlanadi; har bir avtomatik moslikka ishonchlilik balli va inson tomonidan tekshirish holati biriktiriladi; matn tahriri esa qayta yozish emas, provenance qaydi bilan versiyalash tarzida bajariladi. Bu tamoyillar parallel baza keyinchalik semantik, pragmatik va tarjima transformatsiyalarini annotatsiyalashda ma’lumot yo‘qotmasligini ta’minlaydi.



1-rasm. Roman matnini raqamlashtirishdan qidiruvga tayyor parallel bazagacha bo‘lgan ish jarayoni

2. Material va tadqiqot metodlari

(cc) BY

Tadqiqot materiali sifatida “Ulug‘bek xazinasi”ning o‘zbekcha nashri hamda “The Treasure of Ulugbek” inglizcha tarjimasi nazarda tutiladi. O‘zbekcha elektron va bosma nashrlar hajmi va sahifalanishi bo‘yicha farqlanishi mumkin; masalan, 2019-yilgi “Nodirabegim” nashrining bibliografik qaydida 324 bet ko‘rsatilgan [1]. Inglizcha nashr bo‘yicha 2026-yilgi bibliografik yozuvlar mavjud [2; 18]. Shu bois korpusga matn kiritishdan oldin har bir manba uchun doc_id, til, muallif, sarlavha, nashriyot, yil, ISBN, nashr turi, fayl nomi, SHA-256 nazorat yig‘indisi va raqamlashtirish sanasi qayd qilinadi. Bu “edition lock” tamoyili bir nashrdan olingan segmentning boshqa nashr sahifalari bilan aralashib ketishining oldini oladi.

Metodologik asos korpus yaratishning klassik bosqichlari, TEI P5 ning segmentatsiya va parallel matnlarni bog‘lash mexanizmlari, zamonaviy sentence alignment va contextual embedding yondashuvlarini birlashtiradi [5-12]. TEI P5 4.11.0 segmentatsiya tamoyillarini hujjatlashtirishni, segmentlarga xml:id berishni hamda parallel matnlar o‘rtasidagi correspondence/alignment aloqalarini explicit link elementlari orqali ifodalashni qo‘llab-quvvatlaydi [5; 6]. Bu imkoniyat relyatsion bazadagi segment va alignment jadvallari bilan to‘g‘ridan-to‘g‘ri moslashtirilishi mumkin.

Raqamlashtirish jarayonida “xom matnni saqlash” tamoyili asosiy qoida sifatida olinadi. OCR natijasi birdaniga “tozalangan yakuniy matn”ga almashtirilmaydi. Aksincha, kamida uch qatlam saqlanadi: image/facsimile, raw_ocr va normalized_text. Korrektura davomida har bir o‘zgarish source_offset, old_value, new_value, reason va editor maydonlari orqali audit jurnaliga yoziladi. Natijada tadqiqotchi normallashtirilgan korpusdan foydalansa ham, istalgan birlikning dastlabki ko‘rinishiga qaytishi mumkin.

1-jadval

Raqamlashtirishda saqlanadigan matn qatlamlari

Qatlam	Mazmuni	Tahrirga munosabat	Asosiy vazifa
Facsimile	Skan yoki sahifa tasviri	O‘zgartirilmaydi	Asliyatni vizual tekshirish, OCR xatosini aniqlash
Raw OCR	OCR dasturi chiqargan matn	Faqat yangi versiya yaratiladi	Mashina xatosini o‘lchash, qayta ishlashni takrorlash
Diplomatik matn	Asliyat imlosi va punktuatsiyasiga maksimal yaqin matn	Manba bilan tekshiriladi	Filologik va tekstologik tahlil
Normalized text	Kodlash, apostrof, bo‘shliq, texnik belgilar bir xillashtirilgan qatlam	Qoidalar asosida normallashtiriladi	Tokenizatsiya, qidiruv, NLP

(cc) BY

Qatlam	Mazmuni	Tahrirga munosabat	Asosiy vazifa
Abstract layer	Lemma, POS, NER, realia, tarjima aloqalari	Stand-off shaklda qo‘shiladi	Ko‘p qatlamli lingvistik tahlil

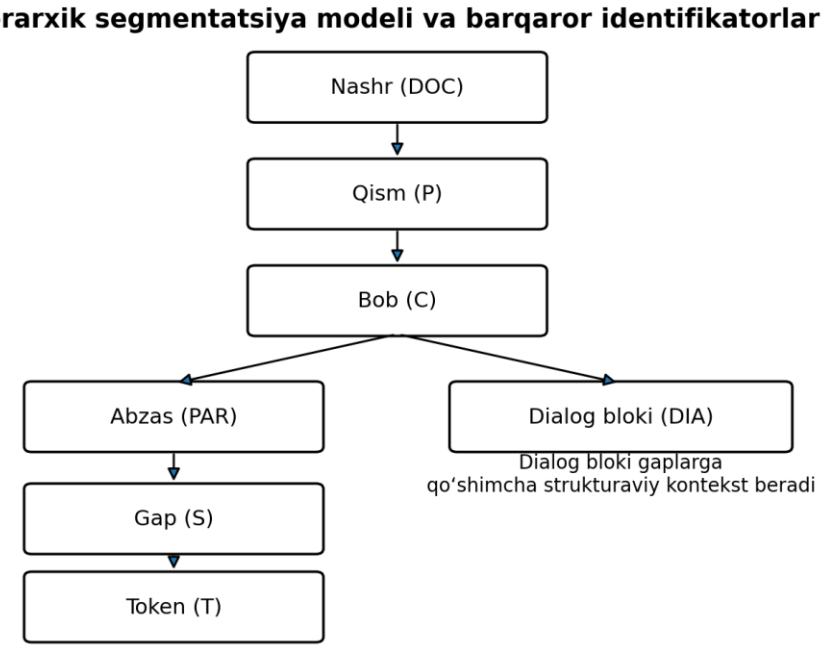
OCR bosqichida o‘zbek lotin yozuvidagi o‘, g‘, apostrof variantlari, ochilish-yopilish qo‘shtirnoqlari, tire va defis, satr oxiridagi ko‘chirishlar, sahifa sarlavhalari va raqamlari alohida nazorat qilinadi. OCRdan keyin Unicode normalizatsiyasi, keraksiz satr uzilishlarini birlashtirish, takroriy bo‘shliqlarni yo‘qotish va tipografik belgilarni standartlashtirish mumkin; ammo leksik shakl, tarixiy birlik yoki muallif uslubini “zamonaviylashtirish” normalizatsiya doirasiga kirmaydi. Bunday o‘zgarish ilmiy matnni buzadi va keyingi tarjima mosliklariga xato signal beradi.

Segmentatsiya qoida asosidagi dastlabki ajratish va inson tekshiruvining kombinatsiyasida bajariladi. Badiiy matnda nuqta, so‘roq va undov belgilari gap chegarasining asosiy ko‘rsatkichlari bo‘lsa-da, qisqartmalar, initsiallar, ko‘p nuqta, muallif izohi, dialog tirelari, qo‘shtirnoq ichidagi gaplar hamda bob sarlavhalari uchun alohida qoidalar zarur. Avtomatik segmentator har bir chegara uchun boundary_confidence qiymatini beradi; past ishonchlilikdagi holatlar reviewer queue ga yuboriladi.

3. “Ulug‘bek xazinasi” matnining segmentatsiya modeli

Parallel korpusning ishonchliligi eng avvalo segmentlarning barqarorligiga bog‘liq. Agar o‘zbekcha matndagi gap identifikatori keyingi korrekturada o‘zgarsa yoki butun korpus qayta raqamlanib ketsa, avval yaratilgan alignment, annotatsiya va havolalar buziladi. Shu sabab segment identifikatori foydalanuvchiga o‘qiladigan ierarxik kod va ichki o‘zgarmas UUID kombinatsiyasida saqlanishi tavsiya etiladi. Masalan, UX-UZ-P01-C03-PAR014-S002 ko‘rinishidagi kod “Ulug‘bek xazinasi - o‘zbekcha - 1-qism - 3-bob - 14-abzas - 2-gap”ni bildiradi. Ma’lumotlar bazasida esa shu birlikka UUID biriktiriladi; matn tahrir qilinsa, UUID saqlanadi, rev_id yangilanadi.

(cc) BY



2-rasm. Segmentlarning ierarxik tashkil etilishi.

2-jadval

Segmentatsiyaning asosiy qoidalari

Holat	Chegara qoidasi	Korpusdagi ifoda	Izoh
Oddiy narrativ gap	Terminal . ? ! dan keyin	S	Qo‘shtirnoq va qavs konteksti hisobga olinadi
Ko‘p nuqta	… faqat sintaktik tugallanish bo‘lsa	S yoki ichki belgi	Avtomatik qoida past confidence bilan belgilanadi
Dialog replikasi	Yangi replika - alohida dialog segmenti	DIA + S	Muallif izohi alohida subsegment bo‘lishi mumkin
Qisqartma/initsial	Nuqta gap chegarasi sifatida olinmaydi	Token ichida	Lug‘at va regex orqali istisno qilinadi
Bob/qism sarlavhasi	Matndan alohida strukturaviy birlik	P/C heading	Gap segmentiga qo‘shilmaydi
Sitatadagi gap	Sitatadagi terminal belgi bo‘yicha	S + quote span	Sitatani alohida annotatsiya qatlamida belgilash mumkin
Satr oxiri ko‘chirish	Gap chegarasi emas	Raw OCR eventi	Normalized qatlamda birlashtiriladi

(cc) BY

Segmentatsiyani faqat gap darajasida cheklash badiiy tarjima uchun yetarli emas. Parallel alignmentda avvalo abzaslar va dialog bloklari tayanch birlik vazifasini bajaradi. Masalan, tarjimon bir o‘zbekcha gapni ikki inglizcha gapga ajratishi yoki ikki gapni bitta inglizcha konstruktsiyada birlashtirishi mumkin. Agar faqat ketma-ket gap raqamlari solishtirilsa, bitta nomuvofiqlikdan keyin butun bob alignmenti siljib ketadi. Shu sabab “coarse-to-fine” usuli qo‘llanadi: avval qism/bob mosligi, so‘ng abzas/dialog bloki, keyin gap, zarur bo‘lsa token darajasi.

Tokenizatsiyada o‘zbekcha agglutinativ tuzilma va inglizcha analitik tuzilma farqi hisobga olinadi. Birinchi bosqichda ortografik tokenlar ajratiladi; keyingi morfologik qatlamda lemma, so‘z turkumi va grammatik belgilar alohida saqlanadi. Tarixiy unvonlar, murakkab atoqli otlar va ko‘p so‘zli realiyalar token darajasida bo‘linishi mumkin, lekin NER/realia qatlamida yagona span sifatida qayta birlashtiriladi. Shu tarzda korpus tokenizatsiya va semantik birlik tushunchalarini aralashtirmaydi.

4. Inglizcha parallel bazani yaratish va alignment modeli

Parallel bazada alignment “tarjima tengligi”ning texnik ifodasidir. Badiiy tarjimada uni faqat ketma-ketlik yoki belgilar soni bo‘yicha aniqlash yetarli emas. Klassik uzunlikka asoslangan yondashuvlar gap uzunligi korrelyatsiyasidan foydalanadi [7], zamonaviy usullar esa ko‘p tilli embeddinglar orqali semantik yaqinlikni hisoblaydi [9-12]. “Ulug‘bek xazinasi” uchun gibrid usul maqsadga muvofiq: strukturaviy anchorlar (qism, bob, abzas, dialog) avval qidiruv maydonini qisqartiradi, keyin semantik model kandidatlarni reytinglaydi, yakuniy murakkab holatlar ekspert tomonidan tasdiqlanadi.

Har bir o‘zbekcha segment u_i va inglizcha kandidat e_j uchun dastlabki moslik balli quyidagicha hisoblanishi mumkin:

$$Score(u_i,e_j) = \alpha \cdot cos(v_{ui}, v_{ej}) + (1-\alpha) \cdot L(u_i,e_j) - \beta \cdot Pboundary$$

bu yerda $\cos(v_{ui}, v_{ej})$ - ko‘p tilli embeddinglar orasidagi kosinus o‘xshashligi; L - uzunlik mosligi; Pboundary - bob, abzas yoki dialog chegarasini kesib o‘tgan kandidat uchun jarima; α va β esa gold-sample asosida kalibrlanadigan koeffitsiyentlardir. Ushbu formula yakuniy hukm emas, candidate generation vositasidir. Badiiy tarjimada semantic similarity yuqori bo‘lsa ham, 1:2 yoki 2:1 munosabat mavjud bo‘lishi mumkin; shuning uchun dinamik dasturlash yoki beam search orqali bir necha ketma-ket segment kombinatsiyalari ham baholanadi.

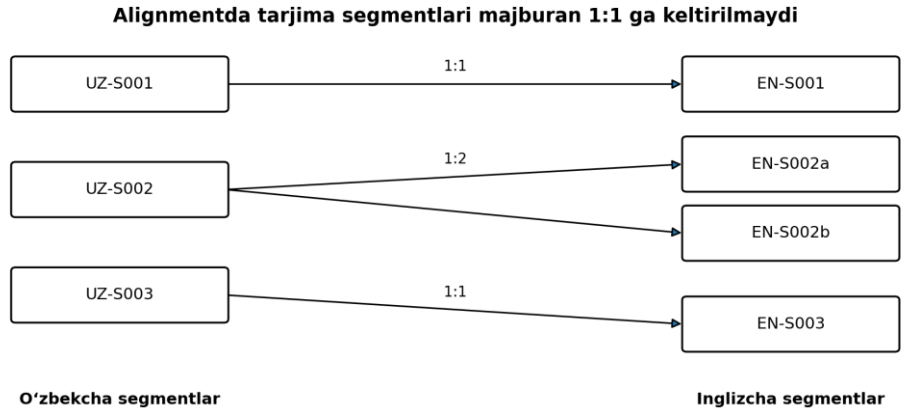
3-jadval

Parallel bazadagi alignment turlari

Tur	Mazmuni	Tipik sabab	Saqlash usuli
1:1	Bitta UZ segment ↔ bitta EN segment	To‘g‘ridan-to‘g‘ri tarjima	src_ids=[1], tgt_ids=[1]

(cc) BY

Tur	Mazmuni	Tipik sabab	Saqlash usuli
1:2 / 1:n	Bitta UZ segment ↔ bir nechta EN segment	Gapni bo‘lish, explicitation	Bitta alignment obyektida ko‘p target
2:1 / n:1	Bir nechta UZ segment ↔ bitta EN segment	Sintaktik birlashtirish, compression	Ko‘p source, bitta target
n:m	Bir nechta segmentlar o‘zaro blokli mos	Qayta tuzish, dialog/narrativ almashuvi	Block alignment, ekspert tekshiruvi
1:0	UZ segment tarjimada tushirilgan	Omission	Null target bilan explicit qayd
0:1	EN tarjimada qo‘shimcha segment	Addition/explicitation	Null source bilan explicit qayd



3-rasm. 1:1 va 1:2 alignment aloqalarini alohida obyekt sifatida saqlash prinsipi.

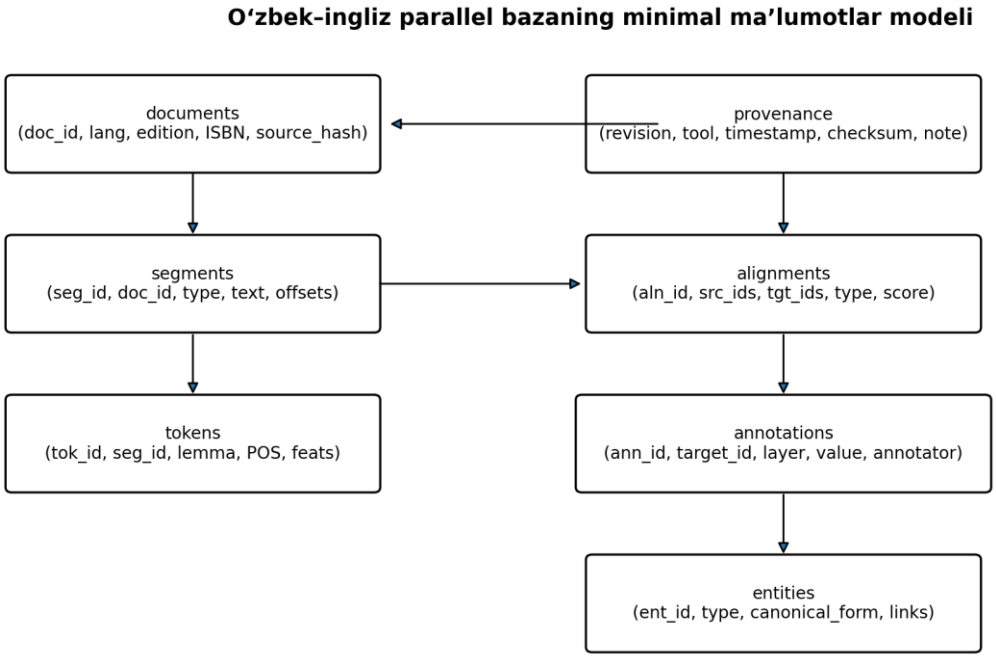
Alignment obyektida faqat ikki matn parchasi emas, balki alignment_type, confidence, method, reviewer_status, reviewer_id va revision maydonlari ham saqlanadi. Masalan, avtomatik tizim 0.93 ishonchlilik bilan 1:1 moslik taklif qilishi mumkin; ekspert uni “accepted” deb belgilaydi. Boshqa holatda 0.61 balli 1:1 kandidat ekspert tomonidan 1:2 ga o‘zgartiriladi. Bu ma’lumotlar keyinchalik modelni qayta o‘qitish, qiyin tarjima hodisalarini topish va annotator qarorlarini tekshirish uchun qimmatli meta-ma’lumot hisoblanadi.

4.1. Ma’lumotlar bazasining strukturasi

Korpusning “parallel baza” qismi faqat ikki ustunli jadvaldan iborat bo‘lmasligi kerak. Ikki ustunli format eksport uchun qulay, lekin ilmiy ma’lumotlar modelida nashr, segment, alignment, token, Abstract va provenance alohida obyektlar sifatida saqlanishi lozim. Bu normalizatsiya bir segmentga bir vaqtning o‘zida bir necha

(cc) BY

annotatsiya qatlamini bog‘lash, yangi tarjima nashrini qo‘shish va oldingi aloqalarni buzmasdan versiyalash imkonini beradi.



4-rasm. Parallel korpusning tavsiya etiladigan relyatsion arxitekturasi.

4-jadval

Asosiy jadvallar va majburiy maydonlar

Jadval	Majburiy maydonlar	Vazifasi
documents	doc_id, lang, title, edition, year, publisher, ISBN, source_hash	Aniq nashr va faylni identifikatsiya qilish
segments	seg_uuid, public_id, doc_id, parent_id, type, order_no, raw_text, norm_text	Har bir strukturaviy/lingvistik segmentni saqlash
alignments	aln_id, src_ids, tgt_ids, type, score, method, status	UZ-EN moslikni ko‘p-kardinal shaklda ifodalash
tokens	tok_id, seg_id, form, lemma, POS, morph, char_start, char_end	Morfologik va token darajadagi ishlov
Abstracts	ann_id, target_id, layer, label, value, annotator, confidence	NER, realia, semantika, pragmatika, translation shift
provenance	rev_id, object_id, action, tool/user, timestamp, before_hash, after_hash	Har bir o‘zgarish tarixini audit qilish

4.2. TEI P5 bilan almashinuv formati

Ma’lumotlar bazasi ichki ishlash uchun qulay bo‘lsa, ilmiy almashinuv va uzoq muddatli saqlash uchun ochiq standart zarur. TEI P5 segmentlar va parallel aloqalarni ifodalash uchun mos keladi: gaplar <s xml:id=…> orqali, alignmentlar esa stand-off <linkGrp>/<link> konstruksiyalari bilan berilishi mumkin [5; 6]. Quyidagi

(cc) BY

soddalashtirilgan fragment mazkur modelning prinsipini ko‘rsatadi; matn mazmuni o‘rniga segment identifikatorlari qoldirilgan.

```
<text xml:lang="uz">
 <s xml:id="UX-UZ-P01-C01-PAR001-S001">…</s>
 <s xml:id="UX-UZ-P01-C01-PAR001-S002">…</s>
</text>
<text xml:lang="en">
 <s xml:id="UX-EN-P01-C01-PAR001-S001">…</s>
 <s xml:id="UX-EN-P01-C01-PAR001-S002A">…</s>
 <s xml:id="UX-EN-P01-C01-PAR001-S002B">…</s>
</text>
<standOff>
 <linkGrp type="translationAlignment">
 <link ana="1:1" target="#UX-UZ-P01-C01-PAR001-S001 #UX-EN-P01-
C01-PAR001-S001"/>
 <link ana="1:2" target="#UX-UZ-P01-C01-PAR001-S002 #UX-EN-P01-
C01-PAR001-S002A #UX-EN-P01-C01-PAR001-S002B"/>
 </linkGrp>
</standOff>
```

Bunday stand-off yondashuvning afzalligi shundaki, asosiy matn annotatsiya bilan ortiqcha yuklanmaydi. Bir xil segmentga morfologik, semantik, nomlangan obyekt, realia va tarjima strategiyasi qatlamlarini mustaqil fayl yoki jadval sifatida bog‘lash mumkin. Bu ayniqsa ko‘p qatlamli semantik bazani keyinchalik qurishda muhimdir.

5. Sifat nazorati va verifikatsiya protokoli

Parallel korpus ilmiy resurs bo‘lishi uchun “matn joylashtirildi” degan holat yetarli emas; har bir bosqich o‘lchanadigan sifat ko‘rsatkichiga ega bo‘lishi kerak. Raqamlashtirishda character error rate (CER) va word error rate (WER), segmentatsiyada boundary precision/recall/F1, alignmentda precision/recall/F1, ekspert qarorlarida esa Cohen kappa yoki Krippendorff alpha kabi inter-annotator agreement ko‘rsatkichlari qo‘llanadi. Ayniqsa badiiy tarjimada alignment uchun faqat avtomatik score bilan cheklanmasdan, tasodifiy va “qiyin holatlar”dan tuzilgan gold-sample zarur.

OCR uchun CER = (S + D + I) / N formulasi qo‘llanadi; bu yerda S - almashtirish, D - tushirib qoldirish, I - ortiqcha belgi, N - etalon matndagi belgilar soni. Segmentatsiya uchun haqiqiy gap chegaralari gold markup bilan solishtiriladi. Alignmentda esa gold-sampledagi to‘g‘ri aloqalar ekspert tomonidan belgilab olinadi; avtomatik tizimning topgan aloqalari shu etalon bilan qiyoslanadi. 1:n va n:m holatlarni

(cc) BY

baholashda alohida “alignment object correctness” mezoni ishlatilishi maqsadga muvofiq: butun src_ids/tgt_ids majmuasi to‘g‘ri topilgandagina aloqa to‘liq to‘g‘ri hisoblanadi.

5-jadval

Sifat nazorati uchun tavsiya etiladigan minimal mezonlar

Bosqich	Metrika	Tavsiya etiladigan maqsad	Tekshiruv usuli
OCR	CER	≤ 1.0% (toza bosma nashr uchun)	Har bobdan reprezentativ sahifalarni qo‘lda etalonlash
OCR	WER	≤ 2.0%	Etalon matn bilan token darajasida qiyoslash
Segmentatsiya	Boundary F1	≥ 0.98	Kamida 1 000 gapli gold-sample
Alignment	Precision	≥ 0.95	Stratifikatsiyalangan qo‘lda tekshirish
Alignment	Recall	≥ 0.95	1:1 va ko‘p-kardinal holatlarni alohida hisoblash
Annotator kelishuvi	Cohen κ / α	≥ 0.80	Ikki mustaqil annotator + adjudication
Provenance	Audit completeness	100%	Har o‘zgarishda revision yozuvi va hash

Jadvaldagi qiymatlar mazkur maqolada erishilgan statistik natija sifatida emas, korpusni ishlab chiqish jarayonida qabul mezoni sifatida taklif etiladi. Bu farq metodologik jihatdan muhim: to‘liq roman korpusi tayyor bo‘lmasdan turib aniqlik foizlarini “natija” sifatida berish ilmiy jihatdan asossiz bo‘ladi. Shuning uchun prototip bosqichida avval gold-sample, threshold va error taxonomy belgilanadi; keyingi maqolada esa real korpus statistikasi aynan shu oldindan belgilangan mezonlar asosida e’lon qilinadi.

Ikki annotator ishtirokidagi tekshiruv ayniqsa quyidagi holatlarda zarur: (a) inglizcha tarjima bitta o‘zbekcha gapni semantik jihatdan ikki gapga bo‘lganda; (b) realiya transliteratsiya qilinib, izoh inglizcha matnning boshqa qismiga ko‘chirilganda; (c) dialog replikasi muallif nutqi bilan qayta strukturallanganda; (d) biror segment tarjimada tushirilganda; (e) tarixiy unvon yoki atoqli otning variant shakli ishlatilganda. Disagreement holatlari uchinchi ekspert adjudication bosqichida hal qilinadi va qaror sababi Abstracts/provenance qatlamida qayd etiladi.

6. Natijalar va muhokama

Tadqiqot natijasida “Ulug‘bek xazinasi” asosida o‘zbek-ingliz parallel bazani qurish uchun oltita asosiy komponentdan iborat formal protokol ishlab chiqildi: (1)

(cc) BY

edition-locked hujjat metadata modeli; (2) raw/diplomatik/normalized qatlamlarga ajratilgan raqamlashtirish tizimi; (3) barqaror ierarxik segment ID sxemasi; (4) ko‘p-kardinal alignment obyektlari; (5) TEI P5 bilan mos stand-off annotatsiya va relyatsion ma’lumotlar modeli; (6) miqdoriy sifat nazorati va provenance auditi. Mazkur komponentlarning bir-biridan mustaqil, ammo ID orqali bog‘lanishi korpusni “yakuniy fayl” emas, evolyutsion ilmiy resurs sifatida yuritish imkonini beradi.

Taklif qilinayotgan modelning badiiy tarjima uchun asosiy afzalligi - 1:1 gap mosligini majburiy norma sifatida qabul qilmasligidir. Tarjima tadqiqotlari “Ulug‘bek xazinasi”dagi madaniy-tarixiy birliklarni ingliz tiliga berishda kontekst, pragmatik funksiya va realia xususiyati muhimligini ko‘rsatadi [16; 18]. Demak, parallel korpusdagi alignment ham sintaktik formalizm bilan cheklanmasligi kerak. Masalan, bir realia asl matnda bitta leksik birlik bilan ifodalanib, tarjimada tavsifiy konstruktsiya orqali berilsa, token soni va gap uzunligi o‘zgaradi; semantik munosabat esa saqlanadi. Gibrid alignment modeli aynan shu tafovutni strukturaviy anchor va semantik yaqinlikni birlashtirish orqali qoplaydi.

Ikkinchi muhim natija - nashr versiyasini korpus obyektining birinchi darajali xususiyati sifatida kiritishdir. Badiiy asarlarning qayta nashrlarida sahifalanish, imlo, punktuatsiya va ayrim tahrirlar farq qilishi mumkin. Inglizcha tarjimaning ham turli bibliografik qaydlari uchrashi sabab “English text” degan umumiy yorliq yetarli emas. documents jadvalidagi edition, ISBN, publisher, source_hash va source_uri kabi maydonlar keyingi ilmiy xulosaning qaysi matn versiyasiga tegishli ekanini aniq ko‘rsatadi. Bu tamoyil korpusning FAIR - topiluvchanlik, kirish, interoperabellik va qayta foydalanish - xususiyatlariga ham xizmat qiladi.

Uchinchi natija - parallel bazani keyingi ko‘p qatlamli semantik bazaga tayyorlashdir. segments va alignments qatlamlari tayyor bo‘lgach, ayni segmentlarga lemma/POS, NER, tarixiy shaxs, joy nomi, mansab/unvon, ilmiy termin, diniy birlik, realia, metafora, pragmatik funksiya va tarjima transformatsiyasi kabi annotatsiyalar stand-off tarzida qo‘shilishi mumkin. Masalan, bir tarixiy unvonning o‘zbekcha shakli, inglizcha ekvivalenti, transliteratsiya varianti va kontekstual izohi bitta entity/realia obyektiga bog‘lanadi. Bu esa oddiy parallel konkordansdan semantik qidiruvga o‘tish imkonini beradi.

Modelning cheklovi shundaki, to‘liq korpus statistikasi, real OCR aniqligi, 1:1 va 1:n nisbatlari hamda annotatorlararo kelishuv ko‘rsatkichlarini nashrning aniq elektron nusxalari va yakuniy inglizcha matn korpusga kiritilmasdan turib ishonchli hisoblab bo‘lmaydi. Shu sabab mazkur maqola metodik-prototip bosqichni ilmiy natija sifatida beradi. Keyingi empirik bosqichda to‘liq matn ikki tomonlama segmentlanadi, stratifikatsiyalangan gold-sample yaratiladi, alignment algoritmlari (uzunlik, embedding, gibrid) qiyosiy baholanadi va xatolar klassifikatsiyasi chiqariladi. Bu

(cc) BY

ketma-ketlik konseptual va empirik natijalarni aralashtirib yubormaslikka xizmat qiladi.

7. Amaliy joriy etish algoritmi

Taklif etilgan modelni amalda joriy etish uchun quyidagi ish ketma-ketligi tavsiya etiladi. Birinchi bosqichda o‘zbekcha va inglizcha nashrlarning huquqiy va bibliografik jihatdan aniq nusxalari tanlanib, metadata va checksum bilan “muzlatiladi”. Ikkinchi bosqichda matnlar OCR yoki born-digital formatdan olinib, raw va normalized qatlamlarga ajratiladi. Uchinchi bosqichda qism-bob-abzas-dialog-gap ierarxiyasi va public_id lar yaratiladi. To‘rtinchi bosqichda bob va abzas anchorlari asosida alignment kandidatlari shakllantiriladi; uzunlik va multilingual embedding ballari birlashtiriladi. Beshinchi bosqichda past ishonchlilikdagi, 1:n/n:1/n:m va null alignment holatlari ekspert tekshiruviga yuboriladi. Oltinchi bosqichda gold-sample asosida sifat ko‘rsatkichlari hisoblanadi va qabul thresholdlari bajarilgandan keyin korpus release versiyasi yaratiladi. Ettinchi bosqichda TEI XML, TSV/CSV va JSONL eksportlari tayyorlanadi; ular konkordans, NLP va keyingi semantik annotatsiya vositalarida ishlatilishi mumkin.

Minimal texnik stek sifatida OCR uchun tilga mos OCR dvigateli, matn normalizatsiyasi va segmentatsiyasi uchun Python, ma’lumotlarni saqlash uchun PostgreSQL yoki SQLite, semantik kandidatlarni topish uchun ko‘p tilli sentence embedding modeli, annotatsiya va tekshiruv uchun esa web-interfeys yoki maxsus corpus Abstract tool qo‘llanishi mumkin. Biroq maqolaning markaziy tamoyili aniq dastur mahsulotiga bog‘lanmaslikdir: identifikatorlar, alignment typology, provenance va TEI eksporti saqlansa, komponentlarni keyinchalik almashtirish mumkin.

Xulosa

“Ulug‘bek xazinasi” romanini raqamlashtirish va inglizcha tarjima bilan parallel bazaga birlashtirish oddiy OCR hamda ikki matnni yonma-yon joylashtirishdan ancha murakkab jarayondir. Badiiy-tarixiy matnning strukturasi, dialoglar, realiyalar, arxaizmlar va tarjima transformatsiyalari segment chegaralarini va 1:1 moslikni muntazam buzadi. Shu sabab mazkur maqolada segmentatsiya va alignment alohida, lekin o‘zaro bog‘langan qatlamlar sifatida modellashtirildi.

Taklif etilgan model besh asosiy tamoyilga tayanadi: birinchidan, korpusga kiritiladigan nashr aniq metadata va checksum bilan identifikatsiya qilinadi; ikkinchidan, raw, diplomatik va normalized matn qatlamlari ajratiladi; uchinchidan, qism-bob-abzas-gap-token ierarxiyasida barqaror segment identifikatorlari yaratiladi; to‘rtinchidan, 1:1 bilan cheklanmagan ko‘p-kardinal alignment obyektlari saqlanadi; beshinchidan, har bir o‘zgarish va annotatsiya provenance hamda versiyalash mexanizmi orqali qayd etiladi. TEI P5 bilan mos stand-off bog‘lash va relyatsion arxitektura esa korpusni boshqa ilmiy tizimlar bilan interoperabel qiladi.

(cc) BY

Mazkur protokol asosida yaratiladigan o‘zbek-ingliz parallel baza keyingi bosqichda morfologik va leksik-semantik annotatsiya, tarixiy realiyalarni avtomatik aniqlash, atoqli otlar va unvonlarni normalizatsiyalash, tarjima transformatsiyalarini statistik tahlil qilish, parallel konkordans yaratish hamda o‘zbek-ingliz mashina tarjimasi uchun sifatli domen ma’lumotini tayyorlashga xizmat qilishi mumkin. Eng muhimi, korpusdagi har bir ilmiy xulosa aniq segment, aniq nashr va aniq revision bilan qayta tekshiriladigan bo‘ladi. Bu “Ulug‘bek xazinasi” singari milliy adabiy merosni raqamli filologiya va kompyuter lingvistikasi doirasida tizimli tadqiq qilish uchun barqaror metodologik asos yaratadi.

Foydalanilgan adabiyotlar

1. Yoqubov O. Ulug‘bek xazinasi: roman. - Toshkent: Nodirabegim nashriyoti, 2019. - 324 b.

2. Yakubov A. The Treasure of Ulugbek / transl. A. Ruziev, D. Rustamov; ed. M. Pierro. - Amazon Digital Services LLC - KDP, 2026. - 166 p. - ISBN 9798195227470.

3. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. - Cambridge: Cambridge University Press, 2012.

4. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. - Cambridge: Cambridge University Press, 1998.

5. TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.11.0. - 2026. Section 17: Linking, Segmentation, and Alignment.

6. TEI Consortium. TEI P5: <segmentation> element: declaration of segmentation principles. Version 4.11.0. - 2026.

7. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. - 1993. - Vol. 19(1). - P. 75-102.

8. Baker M. Corpus Linguistics and Translation Studies: Implications and Applications // Text and Technology: In Honour of John Sinclair / ed. M. Baker, G. Francis, E. Tognini-Bonelli. - Amsterdam: John Benjamins, 1993. - P. 233-250.

9. Jalili Sabet M., Dufter P., Yvon F., Schütze H. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings // Findings of EMNLP 2020. - 2020. - P. 1627-1643. DOI: 10.18653/v1/2020.findings-emnlp.147.

10. Dou Z.-Y., Neubig G. Word Alignment by Fine-tuning Embeddings on Parallel Corpora // Proceedings of EACL 2021. - 2021. - P. 2112-2128.

11. Steingrimsson S., Loftsson H., Way A. SentAlign: Accurate and Scalable Sentence Alignment // Proceedings of EMNLP 2023: System Demonstrations. - 2023. - P. 256-263. DOI: 10.18653/v1/2023.emnlp-demo.22.

(cc) BY

12. Signoroni E., Rychlý P. Evaluating Sentence Alignment Methods in a Low-Resource Setting: An English-Yorùbá Study Case // LoResMT 2023. - 2023.

13. Steingrimsson S. A Sentence Alignment Approach to Document Alignment and Multi-faceted Filtering for Curating Parallel Sentence Pairs from Web-crawled Data // WMT 2023. - 2023. - P. 366-374. DOI: 10.18653/v1/2023.wmt-1.38.

14. Xolmanova Z. Kompyuter lingvistikasi. - Toshkent: Mehr-nuri, 2024. - 232 b. - UDK 81’322(075.8).

15. Rahimov Z. A. Chronotope Image in the Novel “Ulugbek Treasure” by Odil Yakubov // International Journal on Integrated Education. - 2020. - Vol. 3(11). - P. 183-186. DOI: 10.31149/ijie.v3i11.882.

16. Ro‘ziyeva M. Linguistic Problems of Translating Artistic Works from Uzbek to English (in the Example of the Novel “Ulugbek’s Treasure” by Odil Yakubov) // Journal of Academic Research and Trends in Educational Sciences. - 2024. - P. 411-414.

17. Nasirov A. Odil Yoqubovning “Ulug‘bek xazinasi” romanida badiiy obraz yaratish // Uluslararası Türk Lehçe Araştırmaları Dergisi (TÜRKLAD). - 2020. - Vol. 4(2).

18. Ruziev A. Pragmatic Equivalence in Translating Culture-Bound Symbols: The Case of “Xumo Qushi” in Odil Yoqubov’s “Ulug‘bek xazinasi” // International Journal of Literature and Language Studies. - 2026. - Vol. 5(7). Zenodo record 21236656.

19. Chen Y., Eisele A. MultiUN v2: UN Documents with Multilingual Alignments // Proceedings of LREC 2012. - 2012.

20. Pan L., Hang C.-W., Qi H., Shah A., Potdar S., Yu M. Multilingual BERT Post-Pretraining Alignment // NAACL 2021. - 2021. - P. 210-219.

(cc) BY