O‘ZBEK-INGLIZ PARALLEL KORPUSIDA TEGLASH, BILINGVAL KODLASH, SEGMENTLASH, POS TEGLASH VA INDEKSLASH MASALASI

Amirkulov Ma’rufjon Alikulovich,
2-kurs tayanch doktorant
amirkulov.edu01@gmail.com
ToshDO‘TAU

Annotatsıya: Ushbu maqolada o‘zbek-ingliz parallel korpusini yaratish jarayonida muhim bo‘lgan teglash, bilingval kodlash, segmentlash, so‘z turkumlarini avtomatik belgilash (POS-teglash) hamda indekslash masalalari tahlil qilinadi. Korpusni qayta ishlashda ishlatiladigan asosiy formatlar – CoNLL-U, XML, TEI va JSON’ning texnik va lingvistik imkoniyatlari qiyosiy ko‘rib chiqiladi. Har bir formatning afzalliklari, qo‘llanish sohasi va ularning parallel korpusdagi funksional ahamiyati izohlanadi. Tadqiqotda korpusni teglash va kodlashda Universal Dependencies, TEI P5 hamda XML strukturalarining birlashtirilgan modeli tavsiya etiladi. Shuningdek, segmentlash va indekslash jarayonlarining korpusda qidiruv, statistik tahlil va avtomatik tarjima tizimlaridagi roli ilmiy asosda yoritiladi. Maqola natijalari o‘zbek tilining kompyuter lingvistikasi sohasida zamonaviy xalqaro standartlarga integratsiyalashuviga xizmat qiladi.

Kalit so‘zlar: parallel korpus, teglash, segmentlash, POS-teglash, indekslash, CoNLL-U, XML, TEI, JSON, bilingval kodlash, Universal Dependencies.

Abstract: *This article analyzes key processes involved in the development of an Uzbek–English parallel corpus, including annotation, bilingual encoding, segmentation, automatic part-of-speech tagging (POS tagging), and indexing. The main formats used in corpus processing – CoNLL-U, XML, TEI, and JSON are comparatively examined in terms of their technical and linguistic capabilities. The advantages, application domains, and functional significance of each format within a parallel corpus are explained. The study proposes an integrated model that combines Universal Dependencies, TEI P5, and XML structures for corpus*

annotation and encoding. Furthermore, the role of segmentation and indexing in corpus search, statistical analysis, and machine translation systems is discussed on a scientific basis. The results of the study contribute to the integration of the Uzbek language into modern international standards in the field of computational linguistics.

Keywords: parallel corpus, annotation, segmentation, POS tagging, indexing, CoNLL-U, XML, TEI, JSON, bilingual encoding, Universal Dependencies

KIRISH

So‘nggi yillarda tabiiy tilni qayta ishlash (NLP) texnologiyalarining tez rivojlanishi natijasida ko‘p tilli korpuslar, xususan parallel korpuslarga bo‘lgan ehtiyoj keskin ortdi. Bunday korpuslar tarjima, mashinaviy o‘rganish, sentiment tahlili, semantik izlash va lingvistik tadqiqotlar uchun asosiy resurs sifatida xizmat qiladi. O‘zbek tili esa bu borada hali ham past resursli til sifatida ko‘rilmoqda, bu esa o‘zbek tilida avtomatik tahlil va tarjima tizimlarini yaratishda sezilarli to‘siqlar keltirib chiqarmoqda.

Shu nuqtayi nazardan, o‘zbek-ingliz parallel korpusini yaratish faqat til juftligini moslashtirish emas, balki uni to‘liq lingvistik teglash, bilingval kodlash, segmentlash, so‘z turkumlarini avtomatik belgilash (POS tagging) va indekslash orqali kompyuter tahliliga tayyorlashni ham talab etadi. Ushbu bosqichlarning har biri korpusning aniqligi, izchilligi va qayta ishlanish tezligiga bevosita ta’sir qiladi.

Teglash va indekslash uchun ishlatiladigan formatlar (masalan, CoNLL-U, XML, TEI va JSON) o‘z tuzilmasi, texnik moslashuvchanligi va dasturiy integratsiya imkoniyatlari bilan bir-biridan farq qiladi. CoNLL-U soddaligi va mashinaviy qayta ishlash qulayligi bilan ajralib tursa, TEI va XML formatlari kengaytirilgan struktura va boy meta-ma’lumotlar tizimi bilan ilmiy korpuslar uchun ideal hisoblanadi. JSON esa veb-ilovalar va zamonaviy API tizimlariga integratsiya qilishda samarali yechim sifatida ko‘riladi.

Mazkur maqolada ushbu formatlarning texnik, nazariy va amaliy jihatlari chuqur tahlil qilinib, ularning o‘zbek-ingliz parallel korpusida qo‘llanish mexanizmlari, segmentlash va teglash algoritmlari hamda bilingval moslashtirish (alignment) tamoyillari asoslab beriladi.

Teglash, bilingval kodlash, segmentlash, pos teglash va indekslash

O‘zbek-ingliz parallel korpusini yaratishda matnlarni teglash, bilingval kodlash, segmentlash, so‘z turkumlarini avtomatik belgilash (PoS teglash) va ma’lumotlarni indekslash juda muhim bosqichlar hisoblanadi. Ushbu bosqichlar har biri o‘ziga xos dasturiy-uslubiy yechimlar hamda ilmiy yondashuvni talab qiladi[1:35-46]. Quyida har bir masala alohida texnik va nazariy jihatdan ko‘rib chiqilib, ularning amal qilish tamoyillari, formatlari va algoritmlari ilmiy asosda yoritiladi.

Teglash va indekslash formatlari: CoNLL-U, XML, JSON, TEI

Parallel korpusda matnlarni teglash hamda keyinchalik qidiruv indekslarini shakllantirish uchun mos ma’lumot formati tanlash muhim ahamiyatga ega. Hozirgi vaqtda korpus lingvistikasida keng qo‘llaniladigan bir necha formatlar mavjud: CoNLL-U, XML, JSON va TEI. Har bir formatning o‘z afzallik va cheklovlari bo‘lib, tanlangan format korpusning funksionalligiga va qayta ishlash jarayonlariga ta’sir ko‘rsatadi.

CoNLL-U formati

CoNLL-U formati Universal Dependencies ramkasida keng qo‘llaniladigan, 10 ta ustundan iborat sodda tekst formati bo‘lib, har bir token uchun alohida qatorda lingvistik ma’lumotlar beriladi. Jumladan, ID (so‘z tartib raqami), FORM (asl so‘z shakli), LEMMA (so‘zning lemmatik shakli), UPOS (umumiy so‘z turkumi turi), XPOS (tilga xos so‘z turkumi), FEATS (morfologik xususiyatlar ro‘yxati), HEAD (bog‘lanish grafigida bosh so‘z IDsi), DEPREL (sintaktik bog‘lanish turi), DEPS (qo‘shimcha bog‘lanishlar) va MISC (boshqa ixtiyoriy belgi) kabi maydonlar

mavjud[2:1]. CoNLL-U formatining afzalligi – sodda va mashinada oson qayta ishlanishidir: fayl UTF-8 kodida oddiy matn ko‘rinishida bo‘lib, qatordagi ustunlar bir-biridan tab bilan ajratiladi[3: 254-265]. Gap chegaralari bo‘sh qator bilan ajratiladi. CoNLL-U formati ayniqsa sintaktik daraxtlar va morfologik teglangan korpuslarni saqlash uchun qulay bo‘lib, uning standarti turli tillar uchun yagona qat’iy formatni ta’minlaydi[4:1-7]. Misol uchun, quyidagi 1-jadvalda satr bir so‘zning CoNLL-U ko‘rinishdagi teglangan ma’lumotini ifodalashi mumkin :

1-jadval. So‘zning CoNLL-U ko‘rinishdagi teglangan ma’lumoti.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1-gap: "Men kitobni o‘qidim."									
1	Men	men	PRON	P	Person=1|Number=Sing|Case=Nom	3	nsubj	-	-
2	kitobni	kitob	NOUN	N	Case=Acc|Number=Sing	3	obj	-	-
3	o‘qidim	o‘qimoq	VERB	VB	Tense=Past|Person=1|Number=Sing	0	root	-	-
2-gap: "Talaba universitetga bordi."									
1	Talaba	talaba	NOUN	N	Case=Nom|Number=Sing	3	nsubj	-	-
2	universitetga	universitet	NOUN	N	Case=Dat|Number=Sing	3	obl	-	-
3	bordi	bormoq	VERB	VB	Tense=Past|Person=3|Number=Sing	0	root	-	-
3-gap: "Biz bugun imtihondan o‘tdik."									
1	Biz	biz	PRON	P	Person=1|Number=Plur|Case=Nom	4	nsubj	-	-
2	bugun	bugun	ADV	RR	-	4	advmod	-	-
3	imtihondan	imtihon	NOUN	N	Case=Abl|Number=Sing	4	obl	-	-
4	o‘tdik	o‘tmoq	VERB	VB	Tense=Past|Person=1|Number=Plur	0	root	-	-
4-gap: "Professor talabalarni maqtadi."									
1	Professor	professor	NOUN	N	Case=Nom|Number=Sing	3	nsubj	-	-
2	talabalarni	talaba	NOUN	N	Case=Acc|Number=Plur	3	obj	-	-
3	maqtadi	maqtamoq	VERB	VB	Tense=Past|Person=3|Number=Sing	0	root	-	-

XML (Extensible Markup Language)

XML (Extensible Markup Language) – ierarxik tuzilishga ega teglash tili bo‘lib, matnni va uning annotatsiyalarini daraxtsimon ko‘rinishda saqlashga imkon beradi[5:11-20]. XML o‘zbek-ingliz parallel korpusini kodlashda markaziy ahamiyatga ega format sifatida tavsiya etiladi. Sababi, XML korpusni veb orqali ishga tushirish va almashish uchun qulay bir xil kodlashni ta’minlaydi. Har bir matnni yagona sxema bo‘yicha XML’da teglash orqali, barcha hujjatlar uchun

umumiy *“freymvork”* yaratiladi va kelgusida formatinga oid muammolar yuzaga kelmaydi. XML hujjatda odatda ikki asosiy qism farqlanadi: head (bosh) va body (tana) qism[6: 4030-4035].

Bosh qismda matnning muallifi, tarjimoni, yaratilgan davri, uslubi, mavzusi, janri kabi bibliografik yoki kontekstual meta-ma’lumotlar kabi ekstralingvistik ma’lumotlari joylashtiriladi. Tana qism esa matnning o‘zi va lingvistik teglarini o‘z ichiga oladi. Body ichida matn strukturasi (*paragraf, gap, so‘z*) teglar yordamida belgilab chiqiladi. O‘zbek va ingliz tilidagi matnlar alohida <text> bloklarida saqlanib, har birining <text head> (metadata) va <text body> qismlari bo‘ladi. Quyida XML formatidagi bir kichik namuna ko‘rib chiqamiz:

```
<TEXT>
   <TEXT_HEAD>
     <AUTHOR>O'.Yoqubov</AUTHOR>
     <TRANSLATOR>J.Smith</TRANSLATOR>
     <TITLE>Erkinlik</TITLE>
     <YEAR>1980</YEAR>
     <GENRE>Badiiy</GENRE>
   </TEXT_HEAD>
   <TEXT_BODY>
     <p id="1">
       <s id="1">
         <w pos="NOUN">Erkinlik</w>
         <w pos="VERB">qadrlidir</w>
         <pc>.</pc>
       </s>
     </p>
   </TEXT_BODY>
 </TEXT>
```

Ushbu parchada <TEXT_HEAD> qismida matnning muallifi, tarjimoni, sarlavhasi, yili va janri kabi meta-ma’lumotlar ko‘rsatilgan. <TEXT_BODY> qismida esa haqiqiy matn paragraf (<p>) va gap (<s>) darajasida segmentlangan, so‘zlar esa <w> tegi bilan o‘ralgan. Har bir <w> tegida pos atributi orqali so‘z turkumi ko‘rsatilgan. Shuningdek, <pc> tegi (punctuation) bilan tinish belgisi (nuqta) belgilangan. Bunday XML struktura matnning tarkibiy tuzilishini va

lingvistik teglanishini yaxlit holda ifodalaydi. Eng muhimi, parallel matnlarni o‘zaro bog‘lash uchun XML’da qulay imkoniyatlar mavjud: masalan, har bir gapga yoki birlikka yagona id raqami berilib, tillararo moslashtirish (alignment) shu IDlar orqali amalga oshiriladi. Korpusdagi ikki til matnlari <link> elementlari yoki atributlari yordamida bog‘lanadi. Masalan, TEI standartida <code><linkGrp></code> elementi yordamida ikki hujjat orasidagi gaplar mosligi ko‘rsatilishi mumkin. Har bir <link> ichida bir nechta segment IDlari juftligi keltirilib, ular o‘rtasidagi tarjima mosligi va kerak bo‘lsa ishonch darajasi (skor) ham beriladi[7:81-85]. Shu tarzda XML nafaqat bir til doirasidagi teglashni, balki parallel korpus bo‘ylab tillararo bog‘lanishni ham ifodalay oladi[8:61-73].

TEI (Text Encoding Initiative)

TEI (Text Encoding Initiative) – bu aslida XMLning aniqroq ixtisoslashgan standarti bo‘lib, matnli korpuslarni teglash uchun xalqaro qoida va tavsiyalar majmuasidir[9:1]. TEI P5 standarti turli janr va formatdagi matnlarni (*badiiy, ilmiy, og‘zaki nutq, she’riy* va hokazo) teglashga mo‘ljallangan element va atributlarni taqdim etadi. TEI doirasida korpus alohida <teiCorpus> elementi yordamida butun bir to‘plam sifatida ko‘riladi, uning ichida har bir matn alohida <TEI> elementi bilan beriladi. Har bir <TEI> matn ichida yuqorida aytilgan HEAD (header) va BODY qismlari bo‘ladi[10:347-361]. TEI formatining afzalligi shundaki, u standartlashtirilgan: jahondagi ko‘plab korpus loyihalari TEI ga amal qilgani bois, turli korpuslarni bir platformada ishlatish yoki bir xil vositalar yordamida tahlil qilish osonlashadi. Masalan, BMT (Birlashgan Millatlar Tashkiloti) parallel korpusi TEI formatida taqdim etilgan bo‘lib, unda har bir hujjat alohida TEI XML fayl sifatida saqlanadi va hujjatlararo tarjima bog‘lanishlari alohida link fayllari orqali ifodalanadi[11:3530-35354]. TEI standarti parallel korpuslarda segmentlarni bog‘lash uchun <linkGrp> va <link> elementlarini taklif qiladi: masalan, quyidagi <link> ko‘rinishi fransuzcha va inglizcha hujjatlarning 1-gaplari o‘zaro mos ekanini

bildiradi. xtargets="1:1;1:1" atributi birinchi hujjatning 1-gapini ikkinchi hujjatning 1-gapiga bog‘lamoqda. TEI juda boy imkoniyatlarga ega bo‘lsa-da, tuzilishi murakkabroq va hajman kattaroq fayllar hosil qiladi; shu bois kichik va o‘rta hajmdagi maxsus korpuslarda sodda XML yoki boshqa formatlardan foydalanish ham mumkin.

JSON (JavaScript Object Notation)

JSON (JavaScript Object Notation) – oxirgi yillarda ommalashgan ma’lumot almashish formati bo‘lib, an’anaviy korpus formatlarida uncha ko‘p uchramaydi, lekin ba’zi yangi platformalar korpus metadata va annotatsiyalarini saqlash uchun JSON’dan foydalanmoqda[12:166-172]. JSON formati odamga ham, mashinaga ham oson o‘qiladi va ierarxik tuzilmani qisqa yozuvda ifodalash imkonini beradi. Masalan, bitta gapning bilingval teglangan ma’lumotini JSON ko‘rinishida quyidagicha saqlash mumkin:

```
{
   "id": 1,
   "source_sentence": "Bu kichik bir test jumlasidir.",
   "target_sentence": "This is a small test sentence.",
   "tokens": [
     {"form": "Bu", "lemma": "bu", "upos": "PRON"},
     {"form": "kichik", "lemma": "kichik", "upos": "ADJ"},
     {"form": "bir", "lemma": "bir", "upos": "NUM"},
     {"form": "test", "lemma": "test", "upos": "NOUN"},
     {"form": "jumlasidir", "lemma": "jumla", "upos": "NOUN"}
   ]
 }
```

Ushbu misolda gap ID raqami, manba (o‘zbekcha) gap matni va inglizcha gap matni, hamda tokenlar ro‘yxati berilgan. Har bir token uchun form (so‘zning matndagi ko‘rinishi), lemma (bazaviy shakli), upos (universal so‘z turkumi) keltirilgan. JSON formatining afzalligi – u veb-ilovalar va dasturlash tillari bilan integratsiya qilishga juda mos. Ammo ayni paytda an’anaviy korpusni teglash vositalari JSONni kam qo‘llaydi. Ko‘pincha ular XML yoki sathma-sath (tabular) formattdagi fayllar bilan ishlashga mo‘ljallangan. Shunga qaramay, agar parallel korpus uchun maxsus veb-

xizmatlar yoki API yaratilsa, JSON formatida ma’lumotlarni uzatish foydali bo‘lishi mumkin. Ilmiy manbalarda JSON korpus formati keng muhokama qilinmagan, chunki ko‘pchilik annotatsiya tizimlari XML yoki tekst (CoNLL kabi) formatlardan foydalanadi[13:1]. Biroq, JSON orqali korpus tuzilishini ifodalash amaliyoti ham asta-sekin paydo bo‘lmoqda; masalan, ba’zi korpus boshqaruv tizimlari (GitHub’dagi Corpus toolkit kabi) JSON parserlarini taklif qilmoqda[14:1].

Yuqoridagi formatlarning har biri parallel korpusni teglash va indekslashda qo‘llanilishi mumkin. Qiyosiy tahlil qilib aytganda, CoNLL-U sintaktik va morfologik teglangan korpuslar uchun standart bo‘lib, qatorma-qator (“vertical format”) yondashuvi tufayli katta hajmdagi matnlarni minimal hajmda saqlaydi. XML/TEI esa ierarxik yondashuv orqali matnning tuzilmasini, meta-ma’lumotlarni va murakkab annotatsiyalarni bir joyda saqlashga xizmat qiladi; biroq, hajman kattaligi va murakkabligi tufayli uni qayta ishlash uchun kuchli vositalar zarur. JSON esa zamonaviy yengil format sifatida dasturiy ta’minot bilan integratsiya qilish uchun qulay, ammo korpus lingvistikasi an’analarida endigina qo‘llanilayotgan formattur. Parallel korpus yaratishda tanlangan format korpusdan kutilayotgan funksionallikka asoslanadi: agar murakkab qidiruv va ko‘p bosqichli teglash rejalashtirilsa, TEI/XML ma’qul; agar soddalik va tezlik ko‘zda tutilsa, CoNLL-U kabi oddiyroq format kifoya qiladi yoki ma’lumotlar bazasi darajasida JSON ishlatilishi mumkin. Eng muhimi, tanlangan format standartlarga mos bo‘lib, kelgusida boshqa tizimlar bilan uyg‘un ishlash imkonini berishi lozim.

XULOSA

O‘zbek-ingliz parallel korpusi nafaqat tarjima tizimlari, balki lingvistik tadqiqotlar, avtomatik teglash, so‘z turkumlarini aniqlash, semantik tahlil va his-tuyg‘ular (sentiment) tahlili kabi ko‘plab amaliy yo‘nalishlar uchun poydevor vazifasini bajaradi. Maqolada ko‘rib chiqilgan teglash, bilingval kodlash,

segmentlash, POS-teglash va indekslash jarayonlari korpusni sifatli tuzish va undan keyingi hisoblash jarayonlarini optimallashtirishda muhim ahamiyat kasb etadi.

Qiyosiy tahlil natijalari shuni ko‘rsatadiki:

- CoNLL-U formati morfologik va sintaktik tahlil uchun eng qulay, ixcham va standartlashgan shaklni ta’minlaydi;
- XML/TEI formatlari esa murakkab tuzilishli, boy meta-ma’lumotli va parallel korpuslarni xalqaro standartda saqlash uchun eng maqbuldir;
- JSON formati esa veb-ilovalar va API integratsiyasi uchun mos bo‘lib, zamonaviy interaktiv korpus tizimlarini yaratishda foydali hisoblanadi.

Shunday qilib, o‘zbek-ingliz parallel korpusini yaratishda ushbu formatlardan birini yoki ularning kombinatsiyasini qo‘llash orqali lingvistik aniqlik, kompyuter samaradorligi va xalqaro standartlarga moslik ta’minlanadi. Kelgusida ushbu yondashuvlar asosida o‘zbek tili uchun yanada kengroq hajmli, ko‘p darajali teglangan va ochiq manbali korpuslarni ishlab chiqish o‘zbek tilining raqamli rivojlanishiga sezilarli hissa qo‘shadi.

Foydalanilgan adabiyotlar ro‘yxati

1. Doval, I. (2017). POS-tagging a bilingual parallel corpus: methods and challenges. *Research in Corpus Linguistics*, 35-46.

2. https://universaldependencies.org/format.html

3. Rosen, A. (2023). The InterCorp Parallel Corpus with a Uniform Annotation for All Languages. *Jazykovedný časopis*, *74*(1), 254-265.

4. Graën, Johannes, Tannon Kew, Anastassia Shaitarova, Martin Volk, Peter Bański, Adrien Barbaresi, Hanno Biber et al. "Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection." (2019), 1-7.

5. Clarke, K. S. (2011). Extensible markup language (XML). *Understanding Information Retrieval Systems. Management, Types, and Standards, 11-20*.

6. Riaposov, A., & Lazarenko, E. (2024, May). Corpus Services: A Framework to Curate XML Corpus Data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4030-4035).

7. Chen, R., & Liao, H. (2011, May). ParaParse: A parallel method for XML parsing. In *2011 IEEE 3rd International Conference on Communication Software and Networks* (pp. 81-85). IEEE.

8. Zeroual, I., & Lakhouaja, A. (2022). MulTed: A multilingual aligned and tagged parallel corpus. *Applied Computing and Informatics*, *18*(1/2), 61-73.

9. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html

10. Boot, P. (2009). Towards a TEI-based encoding scheme for the annotation of parallel texts. *Literary and linguistic computing*, *24*(3), 347-361.

11. Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016, May). The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3530-3534).

12. Roussel, A. (2024, September). Tabular JSON: A Proposal for a Pragmatic Linguistic Data Format. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)* (pp. 166-172).

13. https://format.digitallinguistics.io/

14. https://github.com/patperry/corpus/blob/master/doc/json.md