TECH SCIENCE

ISSN 3030-3702

TEXNIKA FANLARINING DOLZARB MASALALARI

TOPICAL ISSUES OF TECHNICAL SCIENCES

№ 9 (3) 2025

TECHSCIENCE.UZ

№ 9 (3)-2025

TEXNIKA FANLARINING DOLZARB MASALALARI

TOPICAL ISSUES OF TECHNICAL SCIENCES

TOSHKENT-2025

OAK Ro‘yxati
Mazkur jurnal O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestatsiya komissiyasi Rayosatining 2025-yil 8-maydagi 370-son qarori bilan texnika fanlari bo‘yicha ilmiy darajalar yuzasidan dissertatsiyalar asosiy natijalarini chop etish tavsiya etilgan ilmiy nashrlar ro‘yxatiga kiritilgan.

Muassislar: “SCIENCEPROBLEMS TEAM” mas’uliyati cheklangan jamiyati;
Jizzax politexnika insituti.

TECHSCIENCE.UZ- TEXNIKA FANLARINING DOLZARB MASALALARI elektron jurnali 15.09.2023-yilda 130343-sonli guvohnoma bilan davlat ro‘yxatidan o‘tkazilgan.

TAHRIRIYAT MANZILI:
Toshkent shahri, Yakkasaroy tumani, Kichik Beshyog‘och ko‘chasi, 70/10-uy.
Elektron manzil:
scienceproblems.uz@gmail.com

TEXNIKA FANLARINING DOLZARB MASALALARI
3-jild, 9 -son (Oktyabr, 2025). – 63 bet.

MUNDARIJA

Xo'jayev Otabek, Ro‘zmetova Zilola
MA’LUMOTLARNI INTELLEKTUAL TAHLIL QILISH USULLARI: MASHINAVIY O‘QITISH, CHUQUR O‘RGANISH VA BASHORAT MODELLARI TAHLILI 4-7

Matchonov Shohrux, Asatov Timur
MASHINALI O‘QITISH MODELLARI UCHUN BELGILAR FAZOSINI SHAKLLANTIRISH YONDASHUVI 8-13

Elov Botir, Amirkulov Ma’rufjon, Suyunova Malika
O‘ZBEK-INGLIZ PARALLEL KORPUSIDA METAMA’LUMOTLAR VA FORMATLASH MASALASI HAMDA O‘ZBEK TILI UCHUN MAVJUD NLP VOSITALARI..... 14-21

Jurayev Mansurbek, Khashimov Akhmad
DEVELOPING A NEW MODEL OF CLUSTERING NODES WITH THE HELP OF IOT PROTOCOLS..... 22-30

Ортиков Элбек
СОВЕРШЕНСТВОВАНИЕ СИСТЕМ КОНТРОЛЯ РАСХОДА ГАЗА В ПРОМЫШЛЕННЫХ ПЕЧАХ ДЛЯ ПОВЫШЕНИЯ ИХ ЭНЕРГОЭФФЕКТИВНОСТИ 31-36

Ismailova Zumrad, Anarbaev Anvar, Vaxidov Umid
ISSIQXONADA HARORAT-NAMLIK REJIMINI BOSHQARISH 37-41

Eshmamatov Arslonbek
YASSI QUYOSH HAVO KOLLEKTORLARIDA TO‘SIQ-TURBULIZATORLAR YORDAMIDA ISSIQLIK-GIDRODINAMIK JARAYONLARNI JADALLASHTIRISH ILMIY ASOSLARI..... 42-48

Chuyanov Dustmurod, Bo‘riyev Muhriddin, Fayziyev To‘xtamurod
AVTOTRANSPORT TARMOG‘I KORXONALARINI LOYIHALASHDA ZAMONAVIY YONDASHUVLAR..... 49-54

Abdirazakov Akmal, Boyqobilova Mahliyo
GAZ VA GAZKONDENSAT QUDUQLARINI TA’MIRLASH SAMARADORLIGINI OSHIRISH USULLARI TAHLILI..... 55-62

Texnika fanlarining dolzarb masalalari
Topical Issues of Technical Sciences
2025-yil | 3-jild | 9-son
ISSN (onlayn): 3030-3702

O‘ZBEK-INGLIZ PARALLEL KORPUSIDA METAMA’LUMOTLAR VA FORMATLASH MASALASI HAMDA O‘ZBEK TILI UCHUN MAVJUD NLP VOSITALARI

Elov Botir Boltayevich
Texnika fanlari falsafa doktori, dotsent
ToshDO‘TAU
Email: elov@navoiy-uni.uz

Amirkulov Ma’rufjon Alikulovich
tayanch doktorant
ToshDO‘TAU
Email: amirkulov.edu01@gmail.com

Suyunova Malika Odil qizi
tayanch doktorant
ToshDO‘TAU
Email: malikasuyunova0@gmail.com

Annotatsiya. Ushbu maqolada o‘zbek-ingliz parallel korpusini yaratish jarayonida metama’lumotlarni formatlash, TEI va CoNLL-U standartlaridan foydalanish hamda o‘zbek tili uchun mavjud tabiiy tilni qayta ishlash (NLP) vositalarini tahlil qilish masalalari yoritilgan. Parallel korpusning tuzilishi, ma’lumotlar mosligi, sintaktik va morfologik teglash jarayonlari bosqichma-bosqich ko‘rib chiqilgan. Shuningdek, o‘zbek tili uchun ishlab chiqilgan morfologik analizatorlar, lemmatizatorlar va POS-teggerlar imkoniyatlari tahlil qilinib, ularning parallel korpus yaratishdagi, shuningdek, matnni sentiment tahlil qilishdagi amaliy ahamiyati asoslab berilgan. Tadqiqot natijalari parallel korpuslarni yuqori aniqlik bilan kodlash, avtomatik tahlil qilish va qidiruv tizimlariga integratsiya qilishda metodik asos bo‘lib xizmat qiladi.

Kalit so’zlar: parallel korpus, metama’lumot, TEI, CoNLL-U, o‘zbek tili, NLP, morfologik tahlil, teglash, lemmatizatsiya, lingvistik resurs

METADATA AND FORMATTING ISSUES IN THE UZBEK-ENGLISH PARALLEL CORPUS AND EXISTING NLP TOOLS FOR THE UZBEK LANGUAGE

Elov Botir Boltayevich
Doctor of Philosophy in Technical Sciences, Associate Professor
TSUULL

Amirkulov Ma’rufjon Alikulovich
Principal Doctoral Student
TSUULL

Suyunova Malika Odil qizi
Principal Doctoral Student
TSUULL

Annotatsiya. This article explores the issues of metadata formatting, the use of TEI and CoNLL-U standards, and the analysis of existing Natural Language Processing (NLP) tools for the Uzbek language in the process of creating an Uzbek-English parallel corpus. The paper discusses each stage of corpus development, including text alignment, syntactic and morphological annotation, and structural encoding. Furthermore, it evaluates the performance of Uzbek morphological analyzers, lemmatizers, and POS taggers, emphasizing their practical significance in constructing high-quality bilingual corpora. The results of the study provide a methodological basis for accurately encoding, automatically analyzing, and integrating parallel corpora into linguistic search and processing systems.

Kalit so'zlar: parallel corpus, metadata, TEI, CoNLL-U, Uzbek language, NLP, morphological analysis, tagging, lemmatization, linguistic resources.

DOI: https://doi.org/10.47390/ts-v3i9y2025No3

Kirish

So'nggi yillarda raqamli lingvistika va tabiiy tilni qayta ishlash (NLP – Natural Language Processing) yo'nalishlarining jadal rivojlanishi natijasida turli tillar uchun parallel korpuslar yaratish dolzarb ilmiy-amaliy masalaga aylandi. Parallel korpuslar — bu ikki yoki undan ortiq tildagi matnlar o'zaro moslashtirilgan, ya'ni bir-biriga ekvivalent bo'lgan tarjima birliklari asosida tuzilgan ma'lumotlar bazasidir. Bunday korpuslar tarjima jarayonlarini avtomatlashtirish, mashina tarjimasi tizimlarini o'qitish, semantik va sintaktik tahlilni chuqurlashtirish, shuningdek, tilshunoslikda qiyosiy tadqiqotlar olib borishda katta ahamiyat kasb etadi.

O'zbek va ingliz tillari uchun parallel korpus yaratish ishlari so'nggi yillarda faollashgan bo'lsa-da, bu jarayonda ko'plab texnik va lingvistik muammolar, xususan, metama'lumotlarni to'g'ri formatlash, mos kodlash tizimini tanlash, shuningdek, mavjud NLP vositalarini integratsiya qilish masalalari hanuz to'liq hal etilmagan. Parallel korpusning samarali ishlashi uchun har bir til birligi haqida to'liq va standartlashtirilgan metama'lumot (muallif, manba, sana, janr, til darajasi va h.k.) berilishi, shuningdek, bu ma'lumotlar TEI (Text Encoding Initiative) yoki CoNLL-U formatlari kabi xalqaro standartlarga muvofiq kodlanishi zarurdir.

Bundan tashqari, o'zbek tili uchun ishlab chiqilgan morfologik analizatorlar, lemmatizatorlar va POS-teggerlar (so'z turkumlarini avtomatik aniqlovchi dasturlar) imkoniyatlarini baholash ham parallel korpusni samarali tuzishda muhim ahamiyatga ega. Ayni paytda o'zbek tili uchun Stanza, UDPipe, Hamdam kabi bir nechta NLP vositalari mavjud bo'lib, ularning aniqligi, moslashuvchanligi va parallel korpusdagi ishlov darajasini o'rganish tadqiqotning asosiy yo'nalishlaridan biridir.

Shu sababli, ushbu maqolada o'zbek-ingliz parallel korpusida metama'lumotlarni formatlash, TEI va CoNLL-U standartlaridan foydalanish masalalari, shuningdek, o'zbek tili uchun mavjud NLP vositalarining imkoniyatlari va ularning amaliy qo'llanilishi tahlil qilinadi. Mazkur tadqiqot natijalari parallel korpuslarni yuqori aniqlikda yaratish, ularni avtomatik tahlil tizimlariga integratsiya qilish hamda o'zbek tili uchun zamonaviy lingvistik resurslarni rivojlantirishga ilmiy asos bo'lib xizmat qiladi.

Materiallar va usullar

Metama'lumotlar va formatlash: TEI va boshqa standartlar

Parallel korpus yaratishda metama'lumotlarga alohida e'tibor qaratiladi. Yuqorida ekstralingvistik teglash doirasida aytib o'tganimizdek, har bir matn yoki matn bo'lagi haqida qo'shimcha ma'lumotlar yoziladi (*janr, muallif, sana va hokazo*). Bu metama'lumotlar ma'lumotlarini rasmiylashtirish uchun jahon amaliyotida keng qo'llanuvchi standart – TEI (Text Encoding Initiative) formati mavjud. TEI formati matnli ma'lumotlarni annotatsiya qilish

va ular haqida bibliografik hamda tarkibiy metama'lumotlar saqlashning standartlangan XML asosidagi usulidir. TEI formatida har bir hujjat <teiHeader> degan maxsus bo'lim bilan boshlanadi, unda matnning sarlavhasi, muallifi, nashr ma'lumotlari, tarjimon ismi, janr va boshqa tavsiflovchi ma'lumotlar ko'rsatiladi. Keyin <text> bo'limida matnning o'zi keladi. TEI korpuslarida parallellikni ifodalash uchun bir nechta yondashuv mavjud: ba'zida har ikki til matni alohida <text> bo'limlarida berilib, ularning segmentlari o'rtasida <linkGrp> orqali moslik ko'rsatiladi; ba'zan esa in-place alignment degan usul qo'llanib, tarjima matn original matn ichida maxsus teg bilan belgilab ketiladi. Masalan, TEI formatida parallel segmentlarni ko'rsatish uchun quyidagicha tuzilma bo'ladi:

```
<ab type="paragraph" id="p10">
  <seg xml:lang="uz" id="p10uz">...10-paragrafning o'zbekcha matni...</seg>
  <seg xml:lang="en" id="p10en" corresp="p10uz">...Paragraph 10 in English...</seg>
</ab>
```

Bu misolda <ab> bloki bir parallel paragrafni ifodalaydi, uning ichida o'zbekcha <seg> va inglizcha <seg> ketma-ket berilgan, ingliz segmenti corresp atributi orqali qaysi o'zbek segmentga mos ekanini ko'rsatib turibdi. TEI tavsiyalariga binoan, bunday usul korpusni tarjima xotirasi prinsipi asosida kodlashga o'xshaydi va korpusni qayta ishlash, tahlil qilish hamda undan foydalanishni yengillashtiradi. TEI formatining yana bir yutug'i – u kengaytiriluvchan va moslashuvchan: kerak bo'lsa, o'zimizga xos yangi teglar va atributlar qo'shishimiz yoki mavjud modullardan foydalanishimiz mumkin. Korpusning har bir matniga maxsus <translationQuality> tegi qo'shib, tarjima qanchalik sifatli (masalan, 1dan 5gacha shkalada) ekanini belgilash mumkin bo'ladi.

Natijalar

Parallel korpusni saqlash va almashishda yana bir muhim format – CoNLL-U (yoki CoNLL-X) formati haqida yuqorida to'xtaldik. Aslida CoNLL-U formati avvalo sintaktik treebanklar uchun mo'ljallangan bo'lsa-da, parallel korpus tarkibidagi har bir gapga tegishli barcha lingvistik belgilarni ham bir joyda jamlab ko'rsatishga xizmat qilishi mumkin. Masalan, agar parallel korpusning har bir gap juftligi UD tarzida to'liq teglangan bo'lsa, uni CoNLL-U fayliga yozib, qo'shimcha ustunlarda parallel segment identifikatori yoki tarjima ekvivalenti so'zlar haqida izoh berish mumkin. CoNLL-U fayllari matn shaklida bo'lgani uchun ularni yuklash va o'qish oson, shuningdek, ko'pgina korpus tahlil dasturlari va modellar (masalan, SyntaxNet, UDPipe va boshqalar) to'g'ridan-to'g'ri CoNLL-U formatini qabul qila oladi. Misol tariqasida, yuqoridagi o'zbekcha gapning inglizcha tarjimasi *"She read a big book."* bo'lsin. Agar biz parallel korpusni CoNLL-U'da saqlayotgan bo'lsak, ingliz gapini ham xuddi shu faylda quyidagi tarzda berishimiz mumkin:

***1-jadval**. Ingliz tilidagi gapning CoNLL-U formatdagi tahlil natijasi.*

ID	FORM	LEMMA	XPOS	FEATS	HEAD	DEPREL	ALIGNID
1.	She	she	P	Person=3|Number=Sing	2	nsubj	1
2.	read	read	VB	Tense=Past|Person=3 |Number=Sing	0	root	4
3.	a	a	DET	Definite=Ind|PronType=Art	4	-	-
4.	big	big	ADJ	Degree=Pos	5	-	2
5.	book	book	NOUN	Number=Sing	2	obj	3
6.	.	.	PUNCT	-	-	-	-

Bu misolda inglizcha gap uchun AlignId deb nomlangan qo‘shimcha ustunda har bir so‘zning qaysi o‘zbekcha so‘zga mos kelishi raqamlar bilan ko‘rsatilgan (masalan, *“She”* olmoshi o‘zbek gapidagi 1-olmosh *“U”* ga mos, *“big”* so‘zi o‘zbek gapidagi 2-pozitsiyadagi *“katta”* sifatlarga mos va hokazo). Bu kabi qo‘shimcha belgilashlar parallel korpusdan ko‘p tilli lug‘at chiqarmoqchi bo‘lgan tadqiqotchilar uchun foydali bo‘lishi mumkin. Parallel korpuslarning bunday chuqur belgilangan shakllari sentiment tahlili va ABSA modellarini o‘qitish uchun katta imkoniyat yaratadi. Chunki o‘zbek va ingliz tillarida moslashtirilgan matn juftliklari asosida ko‘p tilli yoki transfer-learning asosidagi ABSA modellarini tayyorlash mumkin bo‘ladi. Masalan, ingliz tilida yirik annotatsiyalangan sentiment datasetlar mavjudligi sababli, parallel korpus orqali bu bilimlarni o‘zbek tiliga o‘tkazish (cross-lingual sentiment transfer) natijasida o‘zbek tili uchun ham yuqori sifatli model yaratish imkoniyati yuzaga keladi. Shu tarzda parallel korpuslar nafaqat tarjimashunoslikda, balki ko‘p tilli his-tuyg‘ularni tahlil qilish tizimlarini shakllantirishda ham asosiy resurs sifatida qaralmoqda. Albatta, bunday chuqur darajadagi markup barcha parallel korpuslarda qilinavermaydi, lekin imkoniyat bor joyda va maqsadga muvofiq hollarda qo‘llash mumkin.

Yakuniy bosqichda, parallel korpusning barcha tarkibi – moslashtirilgan va teglangan matnlar, metadata va indeks ma’lumotlari – yagona tizimga birlashtiriladi. Bu tizim odatda korpus konsol yoki korpus portal shaklida bo‘lib, foydalanuvchiga qidiruv, filtr, solishtirish kabi imkoniyatlarni beradi. Masalan, foydalanuvchi korpus portalida biror so‘z yoki iborani o‘zbek tilida kiritib, uning ingliz tilidagi barcha tarjimalarini ko‘rishi yoki aksincha, inglizcha so‘zning o‘zbek tilidagi tarjimalarini misollar bilan izlab topishi mumkin.

Muhokama

Parallel korpus yaratish jarayoni bosqichma-bosqich amalga oshiriladi:

1. *Matnlarni yig‘ish*
2. *Tozalash va tayyorlash*
3. *Metama’lumotlar bilan teglash*
4. *Gaplarni moslashtirish*
5. *XML kodlash (TEI)*
6. *Morfologik teglash va lemmatizatsiya*
7. *Sintaktik teglash (UD usuli)*
8. *Ma’lumotni CoNLL-U formatida saqlash*
9. *Indekslash va qidiruv tizimi yaratish*
10. *Foydalanish va tahlil*

Bugungi kunda ABSA (Aspect-Based Sentiment Analysis) tizimlari foydalanuvchilarning matnlardagi fikrlarini chuqur tahlil qilishda keng qo‘llanmoqda. O‘zbek tili uchun ABSA tadqiqotlarining rivojlanishi, eng avvalo, sifatli annotatsiyalangan ma’lumotlarga bog‘liq. Shu nuqtayi nazardan, parallel korpuslardan foydalanish ABSA modellarining o‘qitilishida muhim ahamiyatga ega. Chunki ingliz tilidagi yirik sentiment datasetlar (masalan, SemEval korpuslari[1]) bilan o‘zbek tilidagi parallel matnlar moslashtirilganda, cross-lingual transfer learning orqali o‘zbekcha ABSA modelini samarali o‘qitish mumkin bo‘ladi. Bunday yondashuv, bir tomondan, o‘zbek tilida mavjud ma’lumotlar kamchiligi muammosini bartaraf etsa, ikkinchi tomondan, milliy til korpuslarini amaliy sentiment tahliliga bog‘lash imkonini yaratadi. Korpus asosidagi bu jarayonlar, o‘z navbatida, ABSA usuliga tayanch bo‘luvchi ma’lumotlar tahlil zanjirini ham shakllantiradi. Chunonchi, yig‘ish–tozalash–teglash bosqichlari ABSA modelida aspektlarni ajratish, aspekt kategoriyalarini aniqlash va polaritetni baholash uchun zarur

semantik va morfologik ma’lumotlarni tayyorlaydi. O‘zbek tili uchun bunday tayyorlangan parallel korpuslar keyinchalik ABSA modelini o‘qitish, sinovdan o‘tkazish va annotatsiyalangan datasetlar yaratish jarayonlarida asosiy lingvistik resurs sifatida foydalanilishi mumkin. Natijada, parallel korpuslarni ABSA metodologiyasi bilan birlashtirish o‘zbek tilidagi ijtimoiy tarmoq sharhlari, yangiliklar matnlari yoki onlayn ta’lim platformalaridagi fikrlarni tahlil qilishda ham amaliy foyda beradi. Ushbu jarayonlar o‘zaro bog‘liq va iterativ tarzda amalga oshiriladi. Korpusni TEI formatida kodlash uning moslashuvchanligini, CoNLL-U formatida saqlash esa avtomatik tahlil va modellar bilan ishlash imkoniyatini kengaytiradi.

O‘zbek tili uchun mavjud NLP vositalari

Parallel korpus yaratishda, ayniqsa teglash va tahlil jarayonlarida, mavjud dasturiy lingvistik vositalardan foydalanish muhim ahamiyatga ega. So‘nggi yillarda o‘zbek tili uchun bir qator tabiiy tilni qayta ishlash (NLP) vositalari yaratilmoqda. Jumladan, *morfologik analizatorlar, lemmatizatorlar, tokenizatorlar, POS-teggerlar* va hatto ba’zi *sintaktik analizatorlar* ishlab chiqildi.

O‘zbek tilining ilk kompyuter lingvistik resurslaridan biri bu Matlatipov va Vetulani (2009) tomonidan Prolog dasturlash tilida yaratilgan morfologik tahlil dasturidir. Ushbu dastur so‘zning ichki tuzilishini tahlil qilib, uning lug‘aviy ma’nosini va grammatik shakllarini ajratib berishga xizmat qilgan. Biroq, u murakkab qo‘shma so‘zlar va ayrim noan’anaviy shakllarni to‘liq qamrab ololmagani sabab, keyingi tadqiqotlarda takomillashtirish ishlari olib borildi. 2022-yilda Elov va hammualliflar yanada kengaytirilgan taglar to‘plami va chuqur morfologik hamda sintaktik tahlil asosida yangi yondashuvni taklif etdilar[2]. Bu ishlar natijasida 2023-yilda UzNatCorpara deb nomlangan gibrid usul: qoidaga va korpusga asoslangan (rule-based) POS-tegger yaratildi. Ushbu tegger o‘zbek tilining xususiyatlariga mos 12 ta asosiy turkum bo‘yicha so‘zlarni teglaydi va qo‘lda teglangan namunalar ustida sinovdan o‘tkazilgan[3].

Shuningdek, o‘zbek tilida so‘zning asosini topish va lug‘aviy shaklini aniqlash yo‘nalishida ham izlanishlar olib borildi. Elov (2024) turli so‘z shakllarini normal ko‘rinishga keltirish usullarini ishlab chiqdi[4], Sharipov va Sobirov (2022) esa lug‘aviy asoslarni aniqlashga oid yondashuvlarni takomillashtirishdi[5]. 2023-yilda esa Elov, Alayev va Hamroyeva tomonidan UzMorphAnalyzer nomli kompleks dastur taqdim etilib, u stemmer, lemmatizator va POS-tagger funksiyalarini birlashtirgan holda o‘zbek matnlarini to‘liq morfologik tahlil qilish imkonini berdi[6]. Bundan tashqari, ochiq kodli Apertium loyihasi doirasida o‘zbek tilining cheksiz holatlar jamlanmasini qamrab oluvchi yo‘nalishli avtomat (FST) asosidagi morfologik analizator ham yaratilgan bo‘lib, u so‘zning barcha mumkin shakllarini generatsiya va tahlil qila oladi. Ushbu Apertium asosidagi analizator kodi internetda ham mavjud va undan turli NLP vazifalari uchun foydalanish mumkin.

O‘zbek tilini tokenlashtirish (ya’ni matnni so‘zlarga ajratish) masalasi ham muhim vositalarni talab qiladi. Qoida tariqasida, o‘zbek tilida so‘zlar oralig‘ida bo‘shliq mavjud bo‘lsa ham, ba’zi holatlarda birikma va qisqartmalarni to‘g‘ri ajratish uchun maxsus tokenizatorlar kerak bo‘ladi. Masalan, “O‘zbekiston Respublikasi” uchta token bo‘lishi lozim, yoki “stol-stul” birikmasini “stol”, “-”, “stul” tarzida ajratish kabi qoidalarni hisobga oladigan dasturiy ta’minot zarur. Bu borada ham ayrim ishlanmalar mavjud; xususan, 2024-yilda Elov va Xudayberganov hamkorligida o‘zbek tili uchun tokenizator ishlab chiqilganligi haqida xabar berilgan[7].

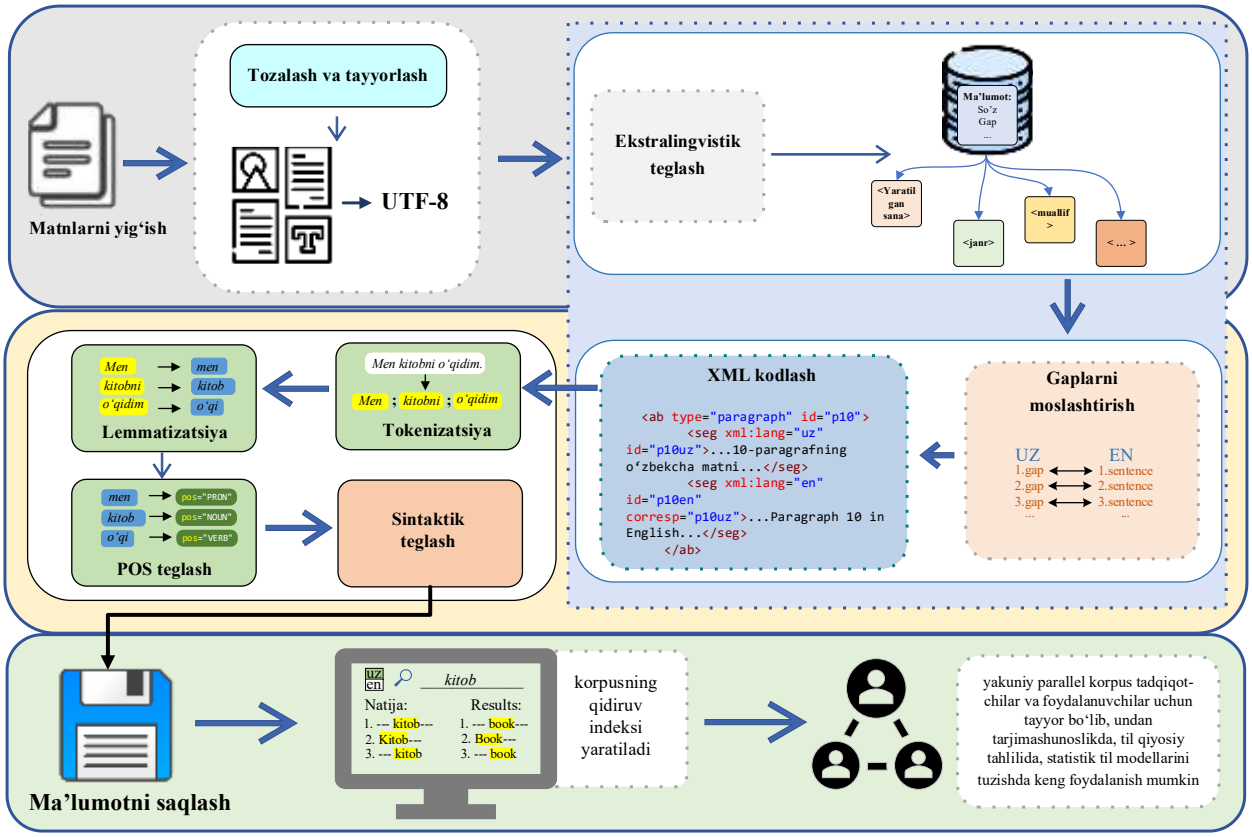
Yuqoridagi vositalarning aksariyatini sinab ko‘rish va foydalanish uchun Alisher Navoiy nomidagi ToshDO‘TAU huzurida yaratilgan maxsus veb-sayt – *“Uznatcorpora”* mavjud.

Uznatcorpora.uz sayti orqali o‘zbek tili morfologik analizatori, lemmatizator, matnlar uchun POS-tegger kabi vositalardan foydalansa bo‘ladi. Masalan, ushbu saytda istalgan o‘zbekcha so‘zni kiritib, uning barcha mumkin bo‘lgan tahliliy shakllarini va grammatik ma’nolarini olish mumkin; yoki matn kiritilganda, har bir so‘zning lug‘aviy shakli va turkumi haqida axborot chiqariladi. Bu vositalardan parallel korpus yaratishda keng foydalanish tavsiya etiladi: matnlarni avtomatik tozalashdan so‘ng tokenizator yordamida so‘zlarga ajratish, morfologik analizator orqali har bir so‘zning grammatik xususiyatlarini aniqlash, lemmatizator yordamida asos shakllarini olish va POS-tegger bilan so‘z turkumlarini belgilash mumkin bo‘ladi. Shuningdek, yuqori darajadagi modellardan ham foydalansa bo‘ladi Albatta, bunday modellardan to‘g‘ridan-to‘g‘ri foydalanish uchun parallel korpuslarning o‘zi kerak bo‘ladi, chunki ular ko‘p miqdordagi ma’lumotda pretren qilish orqali o‘rganadi.

Shu o‘rinda ta’kidlash joizki, mavjud NLP vositalari nafaqat korpus lingvistikasi, balki aspektga asoslangan sentiment tahlili (ABSA) jarayonlari uchun ham bevosita amaliy ahamiyatga ega. ABSA tizimida, ayniqsa, aspektlarni aniqlash (Aspect Term Extraction), aspektlarni toifalash (Aspect Category Detection) va polaritetni baholash (Sentiment Polarity Classification) bosqichlarida so‘zning morfologik shakli, lug‘aviy asosi va turkumi to‘g‘ri aniqlanishi talab etiladi. Shu sababli, o‘zbek tili uchun ishlab chiqilgan lemmatizator, POS-tegger, morfologik analizator kabi vositalar ABSA modelining kirish ma’lumotlarini tozalash va strukturaviy tayyorlashda muhim vosita sifatida qo‘llanishi mumkin.

Masalan, “xizmat sifati yaxshi emas” kabi iboralarni tahlil qilishda “emas” negatsiya elementini, “sifat” so‘zining kategoriyasini va “yaxshi” so‘zining baho komponentini aniqlash aynan NLP vositalari yordamida amalga oshiriladi. Shu tariqa, korpus asosidagi avtomatik teglash tizimlari sentiment tahlilining aniqligini sezilarli darajada oshiradi.

Yuqorida keltirilgan har bir jarayon – matn yig‘ishdan boshlab teglash va vositalardan foydalanishgacha – o‘zaro bog‘liq holda amalga oshiriladi. Quyida parallel korpus yaratish jarayonining umumiy ma’lumot oqimi diagrammasini tasavvur qilish mumkin (1-rasm).



1-rasm. *Parallel korpus yaratish jarayonining umumiy ma’lumot oqimi diagrammasi*

1. Matnlarni yig‘ish – turli manbalardan o‘zbekcha matnlar va ularning inglizcha tarjimalari to‘planadi.

2. Tozalash va tayyorlash – xom matnlar keraksiz elementlardan tozalanib, yagona formatga keltiriladi; gaplarga segmentlanadi.

3. Ekstralingvistik teglash – har bir matnga metadata teglar (janr, manba, sana va h.k.) biriktiriladi.

4. Gaplarni moslashtirish – segmentlangan parallel matnlar avtomatik algoritmlar yordamida gapma-gap align qilinadi; natija qo‘lda tekshirib tasdiqlanadi.

5. XML kodlash – tenglashgan matnlar TEI yoki shunga o‘xshash formatda kodlanib, parallel tuzilma qat’iy belgilab chiqiladi.

6. Morfologik teglash va lemmatizatsiya – korpusdagi har ikki til matnlaridagi so‘zlar avtomatik/qo‘lda teglanadi (tokenlar, lemmalar, POS va morfologik xususiyatlar).

7. Sintaktik teglash – har bir gap uchun sintaktik bog‘lanishlar aniqlanib (masalan, UD usulida) belgilab chiqiladi.

8. Ma’lumotni saqlash – teglangan korpus CoNLL-U kabi formatlarda saqlanishi yoki korpus tizimiga import qilinadi.

9. Indekslash va qidiruv – barcha ma’lumotlar asosida korpusning qidiruv indeksi yaratiladi, bu qidiruv tizimining tezkor va to‘liq ishlashini ta’minlaydi.

10. Tahlil va foydalanish – yakuniy parallel korpus tadqiqotchilar va foydalanuvchilar uchun tayyor bo‘lib, undan tarjimashunoslikda, til qiyosiy tahlilida, statistik til modellarini tuzishda keng foydalanish mumkin.

Yuqoridagi bosqichlar orasida ba’zilari bir vaqtda yoki iterativ ravishda amalga oshirilishi ham mumkin. Masalan, agar korpusni kengaytirish zarur bo‘lsa, yangi matnlarni

qo‘shish uchun yana bir bor yig‘ish-tozalash-moslashtirish bosqichlarini takrorlash talab etiladi. Yoki avtomatik teglashdan keyin qo‘lda tekshirish jarayonida aniqlangan qoidalar asosida tagger dasturini takomillashtirib, qayta teglash ishlarini avtomatlashtirish mumkin bo‘ladi.

Xulosa

Xulosa o‘rnida aytish mumkinki, o‘zbek-ingliz tillari parallel korpusini yaratish – ko‘p bosqichli murakkab jarayon bo‘lib, unda lingvistik va dasturiy tamoyillar chambarchas bog‘liq. Badiiy, rasmiy, yuridik, OAV va ilmiy sohalarni qamrab olgan keng ko‘lamli matnlarni jamlash korpusning qiymatini oshiradi. Ushbu matnlarni puxta tozalab, formal formatga solish keyingi ishlov uchun poydevor yaratadi. Gaplarni moslashtirish algoritmlari – Hunalign, GIZA++ va boshqalar, inson omili bilan uyg‘unlashib qo‘llangandagina yuqori aniqlik beradi. Korpusni TEI formatida kodlash va CoNLL-U kabi strukturalashgan formatlarda saqlash esa uning umrini uzoq qiladi va tadqiqotchilar uchun qulay interfeys taqdim etadi. Nihoyat, o‘zbek tiliga oid yangi yaratilayotgan NLP vositalari va modellarini ishga solish orqali korpusni teglash va tahlil qilish sifati oshadi. Shunday qilib, puxta tashkil etilgan parallel korpus o‘zbek va ingliz tillarini qiyosiy o‘rganish, mashinaviy tarjimani rivojlantirish hamda tilshunoslik nazariyasiga amaliy hissa qo‘shish uchun mustahkam asos bo‘lib xizmat qiladi. Shuningdek, yaratilgan parallel korpus va NLP vositalar kelgusida O‘zbek tilida aspektga asoslangan sentiment tahlil (ABSA) tizimlarini yaratish uchun tayanch infratuzilma bo‘lib xizmat qiladi. Mazkur korpusdagi sintaktik, morfologik va semantik teglash natijalari aspektlarni aniqlash va ularning bahosini aniqlash bosqichlarini avtomatlashtirish imkonini beradi. Shu sababli, o‘zbek tilida korpusga asoslangan sentiment tahlil ishlari ABSA metodologiyasi yordamida yanada chuqurlashadi va tabiiy tilni qayta ishlash bo‘yicha milliy tadqiqotlar uchun yangi ufqlarni ochadi.

Adabiyotlar/Литература/References:

1. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M, & Eryiğit, G. (2016, June). SemEval-2016 Task 5: Aspect-based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 19–30). Association for Computational Linguistics.
2. Mirdjonovna, H. S., & Ilxomovna, A. X. Boltayevich, E. B., (2022, October). Methods for creating a morphological analyzer. In International Conference on Intelligent Human Computer Interaction (pp. 27-38). Cham: Springer Nature Switzerland.
3. Adalı, E., Mirdjonovna, K. S., Xolmo‘minovna, A. O., Yuldashevna, X. Z., Boltayevich, E. B., & Uktamboy O'g'li, X. N. (2023, September). The Problem of Pos Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages). In 2023 8th International Conference on Computer Science and Engineering (UBMK) (pp. 57-62). IEEE.
4. Elov, B., & Xudayberganov, N. (2024). O ‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).
5. Sharipov, M., Mattiev, J., Sobirov, J., & Baltayev, R. (2022). Creating a morphological and syntactic tagged corpus for the Uzbek language. arXiv preprint arXiv:2210.15234.
6. Hamroyeva, S., Alayev, R., Xusainova, Z., & Yodgorov, U., Elov, B. (2023). O ‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and artificial intelligence, 1(3), 117-129.
7. Elov, B., & Xudayberganov, N. (2024). O ‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).

ISSN: 3030-3702 (Online)
САЙТ: https://techscience.uz

TECHSCIENCE.UZ

TEXNIKA FANLARINING DOLZARB MASALALARI

№ 9 (3)-2025

TOPICAL ISSUES OF TECHNICAL SCIENCES

TECHSCIENCE.UZ- TEXNIKA FANLARINING DOLZARB MASALALARI elektron jurnali 15.09.2023-yilda 130346-sonli guvohnoma bilan davlat ro‘yxatidan o‘tkazilgan.
Muassislar: “SCIENCEPROBLEMS TEAM” mas’uliyati cheklangan jamiyati;
Jizzax politexnika insituti.

TAHRIRIYAT MANZILI:
Toshkent shahri, Yakkasaroy tumani, Kichik Beshyog‘och ko‘chasi, 70/10-uy.
Elektron manzil:
scienceproblems.uz@gmail.com