ISSN 2181-922X 97—109

O‘zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari

Botir Elov[1]
Ma’rufjon Amirqulov [2]

Abstrakt

Maqolada kompyuter lingvistikasining asosiy yo‘nalishlaridan biri va ko‘plab masalalari yechilishi lozim bo‘lgan korpus lingvistikasi, monolingual korpuslar hamda parallel korpuslar haqida, bundan tashqari parallel korpus sohasida jahon tajribasiga tayangan holda o‘zbek-ingliz tillarining parallel korpusini yaratish bosqichlari haqida so‘z yuritiladi. O‘zbek-ingliz tillari parallel korpusining dasturiy va lingvistik tamoyillarini asoslash, tanlangan birliklarni lingvistik va ekstralingvistik teglash, parallel korpus tuzish algoritmini ishlab chiqish kabi ustuvor vazifalar haqida ma’lumotlar beriladi. Parallel korpusga qanday ma’lumotlar tanlanishi, ma’lumotlarga qo‘yiladigan talablar va o‘zbek-ingliz parallel korpusining yaratilishi tadqiqotchi va foydalanuvchilarga qanday imkoniyatlar taqdim etishi to‘g‘risida fikrlar yuritiladi. Ushbu jarayonda material tanlash kabi lingvistik va uslubiy muammolar bilan bir qatorda parallel korpusni yaratishdagi dasturiy qiyinchiliklar xususida xulosalar aks etadi.

Kalit so‘zlar: *Korpus, teglash, tenglashtirish, parallel korpus, Pos teglash, XML kodlash*

Kirish

So‘ngi yillarda korpus lingvistikasi jadallik bilan rivojlanib bormoqda. Parallel korpuslarning qo‘llanishi amaliy tilshunoslik tadqiqotlari ko‘lamini yanada kengaytiribgina qolmay, chet tilini o‘rgatish bo‘yicha olib boriladigan tadqiqotlar uchun yangi falsafiy

[1]*Elov Botir Boltayevich* – texnika fanlari bo‘yicha falsafa doktori (PhD), dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi mudiri.
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648

[2]*Amirqulov Ma’rufjon Aliqul o‘g‘li* - Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxassisligi 2-kurs magistranti.
E-pochta: amirkulovmaruf@navoiy-uni.uz
ORCID: 0000-0002-4025-8466
Iqtibos uchun: Elov B., Amirqulov M. 2022. “O‘zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari”. *O‘zbekiston: til va madaniyat. Amaliy filologiya*. 2 (5): 97-109.

qarashlarni shakllantirmoqda [Wen, Wang, Liang, 2005]. Parallel korpus atamasi odatda lingvistik doiralarda bir-birining tarjimasi bo'lgan matnlarga murojaat qilish uchun ishlatiladi. Bir tilli hisoblanmish monolingual korpuslarni butun dunyo bo'ylab keng ko'lamda uchratish mumkin. Ingliz tili monolingual korpusi, ayniqsa, eng katta hajmni egallaydi. Monolingual korpuslarni yaratishdagi muvaffaqiyat bilan taqqoslanganganda, parallel korpuslar hajmi, ayniqsa o'zbeklar ishtirok etgan parallel korpuslar, unchalik ko'p emas va bu o'zaro tillarda axborotni qayta ishlash, til o'rgatish, tadqiqot hamda ikki tilli lug'at tuzish kabi jarayonlarning istiqboliga sezilarli darajada to'sqinlik qiladi. Dunyo kompyuter lingvistikasida oxirgi 30 yil davomida parallel korpuslarni avtomatik tarjima platformasi, tezaurus, elektron lug'at sifatida qo'llash va ularni ilmiy-nazariy jihatdan o'rganish harakati avj oldi. Shu jarayonda o'zbek-ingliz, ingliz-o'zbek parallel korpusini tuzishning dasturiy va lingvistik tamoyillarini asoslash, tanlangan birliklarni lingvistik va ekstralingvistik teglash, parallel korpus tuzish algoritmini ishlab chiqish kabi ustuvor vazifalar morfologiya, sintaksis, tarjimashunoslik sohalarida uzusning formal va noformal registrini aniqlash, faqat tilshunos intuisiyasiga tayanilgan, subyektivlikka moyil tadqiqotlarning ishonchliligi va obyektivligini ta'minlash, yangi avlod korpus lug'atlari va korpus grammatikalarini yaratishga keng imkon yaratadi [Karimov, 2022].

Yuqori sifatli tarjimaga ega parallel korpusni yaratish kompyuter lingvistikasining mashina tarjimasi sohasidagi dolzarb masalalardan biriga aylanib bormoqda. Afsuski, bunday yuqori sifatli parallel korpusni yaratish juda ko'p sabablarga ko'ra qiyin masala hisoblanadi. Sifatli tarjimalarning mualliflik huquqini sotib olish, parallel matnlarni, gaplarni va so'zlarni muqobil tenglashtiruvchi dasturlarni topishdagi qiyinchiliklar, korpusni yaratishdagi yuqori sarf-xarajatlar shular jumlasidandir [Park, Lee, Eo, Seo, Lim, 2021]. Parallel korpus uchun manba tanlashdagi eng ishonchli ma'lumotlar sifatida inson tomonidan bajarilgan tarjimalarni olishimiz mumkin [Hutchins, 2001, 5; Rojo, 2018, 257]. Inson omili orqali tarjima qilingan matnlar orqali korpusning aniqlik darajasi ortadi, mashina tarjimasi uchun esa sifatli ma'lumotlar bazasi sifatida xizmat qiladi hamda foydalanuvchilarga mavjud avtomatik tarjima dasturlar bilan solishtirganda ancha yuqori sifatdagi tarjimaga ega bo'lish imkoniyatini yaratadi. Bu jarayon esa korpusdagi to'g'ri va sifatli materiallar qamrovi ortgani sababli yanada yaxshilanib borishi, parallel korpuslarni yaratish kompyuter lingvistikasidagi nechog'lik

muhim masalalardan biri ekanligini namoyon qiladi.

Asosiy qism

Eng keng tarqalgan fikr va qarashlarga ko‘ra “korpus” (lotinchadan so‘zma-so‘z “tana” ma’nosida) yozma matnlar, ayniqsa muayyan bir yozuvchining muayyan mavzularidagi asarlarini qamrab olgan jamlanma sifatida qaralgan [Saad, 2015]. Korpus lingvistikasini tilshunoslikda foydalanish uchun tuzilgan, matnlarning katta hajmdagi bazasi orqali tilni o‘rganish metodologiyasi sifatida ko‘rish mumkin [Leech, 2010, 103]. Korpus va korpus lingvistikasiga doir dastlabki tadqiqot hamda rivojlanish jarayonlari asosan ingliz tilshunosligi doirasida amalga oshirilgan. 1960-yillarda yaratilgan Braun korpusi birinchi mashinada o‘qiladigan korpus sifatida tarixda iz qoldirgan bo‘lsa, undan so‘ng yaratilgan Lankaster-Oslo/Bergen (LOB) korpusi [Svartvik, 1992, 7] haqida ham shunday deyish mumkin. Korpusning mavjudligi katta hajmdagi lingvistik ma’lumotlarga ega bo‘lishga imkon yaratibgina qolmay, ko‘plab kvantativ va variatsion tadqiqotlar uchun ham dasturulamal bo‘lib xizmat qildi. Dastlabki ikkita korpus (Braun va LOB) bir xil dizayn mezonlari bo‘yicha tuzilgan edi, bu esa bir xil tuzilgan korpuslardagi ingliz tilining ikki xil variantini qiyoslash g‘oyasini olib keldi. Natijada korpus metodologiyasiga tamoman begona bo‘lgan qiyoslash nuqtayi nazari kirib keldi. Biroq, ko‘p yillar davomida korpuslar bir tilli bo‘lib qoldi, bu esa qiyosiy tatqiqotlar imkoniyatini cheklovchi sabab edi. Qiyoslash faqatgina boshqa tillarda yaratilgan bir tilli korpuslar orqali amalga oshirilardi. Bitta korpusda ikki va undan ortiq tillarning parallel ma’lumotlarini topish korpus lingvistikasi rivojlanish bosqichining ilk davrlarida yetishib bo‘lmas jarayon edi. Ko‘p tilli korpuslarni yaratish haqidagi fikrlar yaqin yillarga borib taqaladi. 1990-yillar boshlarida olimlar Stig Yoxansson va Knut Xofland ingliz-norveg tillari parallel korpus (ENPC)i uchun harakatlarni boshlashgan [Johansson, Hofland, 1994, 25]. Yaratilishi mo‘ljallangan korpus ingliz va norveg tillarining asl va tarjima matnlarini o‘z ichiga olishi hamda unda ma’lumotlarni tenglashtirish va qidiruv tizimining yangi texnologiyalari yaratilishini talab qilar edi. Stig Yoxanssonning boshqa loyihalari singari, ingliz-norveg tillari parallel korpusi uning ingliz-shved tillari parallel korpus (ESPC)i mualliflari hisoblanmish shved hamkasblarining yordami bilan yaratilgan edi.

Ma’lumotlarni tenglashtirish uchun foydalaniladigan dasturiy ta’minot Knut Xofland tomonidan yaratilgan bo‘lsa, Jarle Ebel tomonidan parallel konkordans uchun tizim yaratildi. Ingliz-shved

tillari parallel korpusidagi asl va tarjima matnlar gaplar kesimida tenglashtirilgan bo'lib, har bir gap yoki gaplar birligi boshqa tildagi xuddi o'sha gap yoki gaplar birligiga o'ziga xos xususiyatga ega identifikatsiya tegi bilan biriktirilgan edi. Quyidagi misolda ushbu korpusdagi teglangan va tenglashtirilgan gapdan namuna keltiramiz:

(1a) <s id=ABR1.1.1.s326 corresp=ABR1T.1.1.s325>But how come you speak the language so fluently?" </s>(ABR1)

(1b) <s id=ABR1T.1.1.s325 corresp=ABR1.1.1.s326>Men hvordan har det seg at De snakker språket så flytende?" (ABR1T)

Yuqoridagi kabi ikki va ko'p tilli korpuslarning turli jabhalardagi o'rni beqiyos. Qiyosiy tahlil jarayonida ham parallel korpuslarning nechog'lik muhimligi quyida keltiriladigan fikrlar orqali dalillanadi. Qiyosiy tahlil – bu ikki yoki undan ortiq tillarni maqsadli ravishda ularning farqli va o'xshash tomonlarini tizimli taqqoslash hamda tavsiflash demakdir [Johansson, 2007]. Parallel korpuslarning qiyosiy tahlil jarayonidagi ustunliklari xususida olimlar Aijmer va Altenberglar tomonidan tushuntirib berilgan:

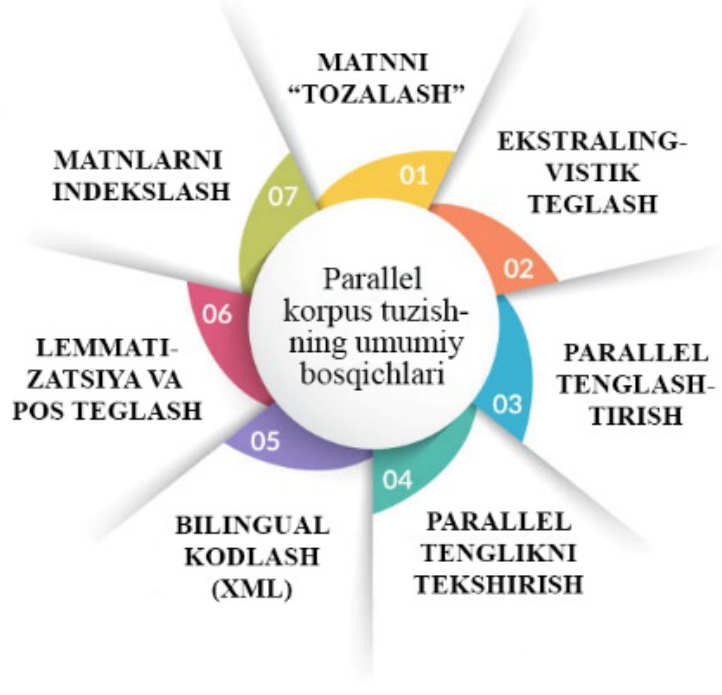
- Ular monolingual korpuslarda erishib bo'lmaydigan ustunlik, ya'ni tillarni taqqolash imkonini beradi;
- Ular bir qator qiyosiy maqsadlarda ishlatilishi va bizning tilga xos tipologik va madaniy farqlar, shuningdek, umumiy o'ziga xos xususiyatlar haqidagi tushunchamizni oshirishi mumkin;
- Ular manba va tarjima, mahalliy va nomahalliy matnlar o'rtasidagi farqlarni o'rganishda ishlatiladi;
- Ular bir qator amaliy dasturlar, masalan, leksikografiya, tilni o'rgatish va tarjima sohasida ishlatilishi mumkin.

Bir tilli korpuslar singari, parallel korpuslar leksika, leksika – grammatika va nutqiy xususiyatlarni o'rganish uchun juda mos keladi. Korpusning teglangan versiyasi esa o'sha grammatik konstruksiyalarni o'rganishni yanada osonlashtiradi.

Korpusni yaratishga doir ko'plab adabiyotlar ma'lum bo'lib, ulardan birining sharti korpus muvozanatlangan bo'lishini taqozo etadi. Bunda turli soha va janrlarga tegishli matnlar korpusdan nisbatan teng hajmni egallashi kerak, ana shunda o'sha tilning bor jozibasini ko'rish mumkin. Biroq, parallel korpuslarni, xususan, o'zbek-ingliz tilllari parallel korpusini tuzishda yuqoridagi talabga amal qilish qiyin. Chunki ushbu tillarning elektron bilingual matnlarini topish anchayin mushkul masala. Hozirda bir necha o'zbek asarlarining ingliz tiliga, ingliz asarlarining o'zbek tiliga o'girilgan variantlarini topish mumkin, ammo bu ham ikki tilning jozibasini to'laqonli ochib berolmaydi va hajman parallel korpus

uchun kam miqdor hisoblanadi. Shunday bo‘lsada, mamlakatimizda korpus lingvistikasi doirasida parallel korpuslarni yaratishdagi ilk qadamlar sifatida o‘zaro yaxshi tarjima qilingan matnlarning elektron shakli qancha miqdorda topilsa, ularning barchasidan unumli foydalanish maqsadga muvofiq bo‘ladi. Demak, o‘zbek-ingliz tili juftligidagi matnlarni internet orqali, masalan, HTML fayl shaklida yoki turli elektron hujjatlar shaklida yig‘ish mumkin. Ushbu holatda keraksiz HTML teglari, chizmalar, jadvallar fayllardan olib tashlanadi va korpusga kiritish uchun faqat matndan iborat shaklga keltiriladi.

Ko‘plab olimlarning tadqiqotlariga binoan, korpusga, xususan o‘zbek-ingliz tillari parallel korpusiga kiritilishidan oldin, har qanday matn quyidagi jarayonlardan o‘tmog‘i lozim:



Matnni “tozalash”. Parallel korpus uchun manba tanlashda ko‘plab matnlar internet tarmog‘i orqali jamlanadi va shu sabab ularning ko‘pi HTML fayllardan tashkil topadi. HTML fayllarda matnlardan tashqari ko‘plab teglar, ortiqcha havolalar mavjud bo‘lishini hisobga olgan holda, daslabki bosqichda parallel matnlar ushbu keraksiz belgilardan holi qilinadi.

Ekstralingvistik teglash. Ushbu bosqichda matnlarga ekstralingvistik teglar biriktiriladi. Masalan, matnning qaysi sohaga tegishli ekanligi, kim tomonidan yozilganligi yoki tarjima qilinganligi,

davri, qaysi janrga tegishli ekanligi ekstralingvistik teglash doirasida amalga oshiriladi.

Parallel tenglashtirish. Parallel matnlar paragraflar va gaplar kesimida tenglashtiriladi. Ushbu bosqichni amalga oshirish uchun ko'plab avtomatik dasturlar yaratilgan bo'lib, ular bir-biridan aniqlik darajasi, ma'lum bir til yoki til oilasiga moslashtirilganligi va yana boshqa xususiyatlariga ko'ra farq qiladi. Jahondagi ko'plab parallel korpuslar Gale va Church [Gale, Church 1993, 75] taklif etganidek, gaplar kesimida tenglashtirilgan. Bunda parallel matnlardagi gaplar uzunligiga qarab bir-biriga tenglashtiriladi.

Parallel tenglikni tekshirish. Yuqorida ta'kidlanganidek parallel matnlarni tenglashtirishda bu jarayonni avtomatik amalga oshiruvchi dasturlardan foydalaniladi. Biroq bu dasturlar 100% aniqlikda matnllarni tenglashtiradi deb hisoblay olmaymiz. Chunki, ingliz tilidagi bitta gap bilan ifodalangan matn bo'lagi o'zbek tilida ikkita gap bilan ifodalanishi mumkin yoki aksincha. Bu holatda avtomatik tenglashtiruvchi dastur parallel matnlarni faqat 1:1 nisbatda tenglashtiradi. Birga ko'p va ko'pga bir tenglashtirish uchun esa inson omili kerak bo'ladi. Shu sabab ushbu bosqichda qandaydir xatolar inson omili orqali tekshiriladi va to'g'rilanadi. LF Aligner, E Align va shu kabi avtomatik tenglashtiruvchi dasturlar yaratilgan bo'lib, ular ingliz tili va boshqa tillar matnlarini gap hamda paragraf kesimida tenglashtiradi.

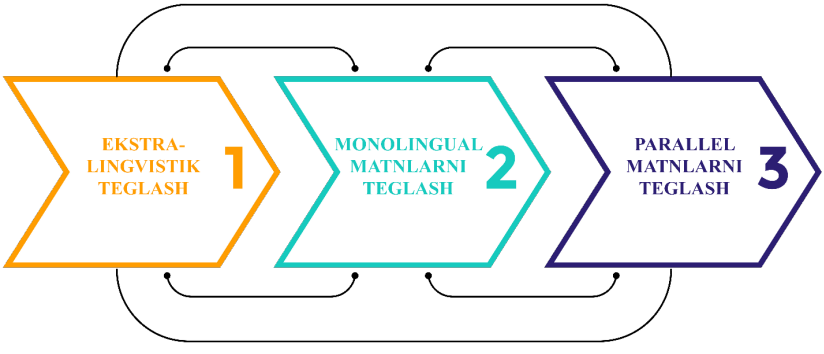
Bilingual kodlash (xml). Ushbu bosqichda tasdiqlangan matnlar XML formatda kodlanadi. Kodlash avtomatik XML kodlovchi dasturlar orqali amalga oshiriladi.

Lemmatizatsiya va pos teglash. Bugungi kunda parallel korpus sifatini oshirish hamda foydalilik ko'lamini oshirish uchun lemmatizatsiya va Pos teglash muhim omil bo'lib xizmat qiladi. Ushbu jarayonlarni amalga oshirish katta mehnat talab qilsada, parallel korpusning ko'plab xususiyatlarga ega bo'lishiga samarali hissa qo'shadi.

Matnlarni indekslash. Oxirgi bosqichda hamma tekshirishlardan o'tgan matnlar indekslanadi. Indekslash jarayonining amalga oshirilishi qidiruv tizimining optimallashishiga yordam beradi. Ya'ni katta matnli ma'lumotlar kichik-kichik qismlarga bo'linib indekslanadi va bu esa qidirilayotgan natijaga erishish vaqtini sezilarli darajada tezlashtiradi.

Parallel korpusni teglash masalasi. Korpus faqatgina ma'lumotlari teglangandan so'nggina foydali bo'lishi mumkin. Masalan, turli sathlarda teglangan parallel korpusgina **tarjima**

xotirasi bo‘lib xizmat qiladi. Parallel korpus uchun eng muhim teglash bu **tenglashtirish**dir. Gaplar parallel korpuslarning eng muhim tenglashtirish mahsuloti hisoblanadi. Monolingual korpuslarga tegishli teglarni parallel korpuslarga ham qo‘llash mumkin. Misol uchun Pos (so‘z turkumi) va sintaktik teglash shular jumlasidandir. Biroq, bu jarayonlarni amalga oshirish katta mehnat va sabrni talab qilishini unutmaslik lozim. Jahon tajribasi hamda bir nechta mashhur parallel korpuslarga tayangan holda korpusni teglashning 3 ta turini aniqladik:



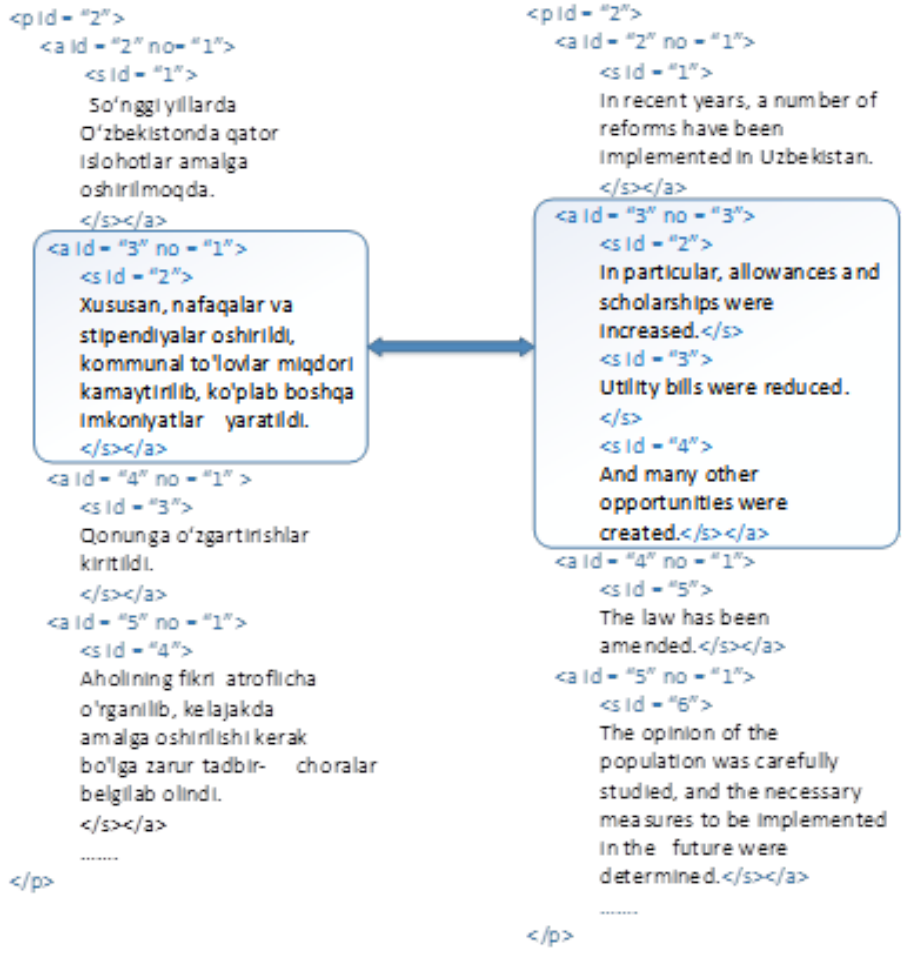
Ekstralingvistik teglash. Ekstralingvistik teglash o‘z ichiga parallel korpusga kiritilgan to‘liq bo‘lgan matnlar haqidagi ekstralingvistik ma’lumotlarni qamrab oladi. Masalan, yuqorida ta’kidlangani kabi matnning muallifi, tarjimoni, sohasi, yozma yoki ovoz shaklida ekanligi, qaysi davrga tegishli ekanligi va sarlavhasi shular jumlasidandir. Parallel korpusdagi ekstralingvistik teglangan ma’lumotlar ko‘plab tadqiqotchilar uchun qo‘l kelishi mumkin. Masalan, tadqiqotchilar faqatgina bir sohaga, bir davrga yoki bir yozuvchiga tegishli bo‘lgan matnlarga qiziqayotgan bo‘lsa ekstralingvistik teglangan ma’lumotlar qidiruvi orqali o‘zi izlayotgan natijaga soniyalar ichida erishishi mumkin [5].

Monolingual matnlarni teglash. Ushbu teglash jarayoni parallel korpusdagi ikki til uchun alohida amalga oshiriladi. Matnlar, paragraf, gap va so‘zlar kesimida annotatsiyalanadi. Bundan tashqari o‘zbek-ingliz tillari parallel korpusida o‘zbek tilidagi so‘zlar ustida ularning lemmasini biriktirish hamda Pos (so‘z turkumlari) teglash jarayoni amalga oshiriladi. Ingliz tilidagi so‘zlar uchun ham shu jarayon qo‘llanadi.

Parallel matnlarni teglash. Bu bosqichda ikki til o‘rtasida “ko‘prik” o‘rnatiladi. Original matn va ularning tarjimalari paragraf va gaplar kesimida tenglashtiriladi. So‘zlar kesimida tenglashtirishni amalga oshirish hozirgi vaqtda qiyin vazifa hisoblanadi. Birinchidan,

so‘z kesimida tenglashtiruvchi dasturlarning yetishmasligi bo‘lsa, ikkinchidan, bu jarayon uchun katta mehnat talab etiladi.

Bilingual kodlash (XML) masalasi. Yaratilgan korpusni internet orqali ishga tushirishga qulay bo‘lishi uchun barcha matnlar bir xil kodlanishi zarur [Baobao, 2004]. Shu sabab, XML ga asoslangan freymvork tanlanishi maqbul tanlovdir. Bunda barcha o‘zbek va ingliz matnlari alohida kodlanadi va ushbu matnlar tilidan qat’iy nazar matnning “bosh”i va “tana”sidan tashkil topadi. Barcha ekstralingvistik ma’lumotlar bosh qismiga, monolingual teglar, lingvistik ma’lumotlar hamda matnning o‘zi tana qismiga joylashtiriladi. Parallel matnlarni tenglashtirishlar esa ikki tilning tana qismidagi id raqamlari orqali amalga oshiriladi. Quyida o‘zbek-ingliz tillaridagi matnning bir qismi tenglashtirilishining taxminiy ko‘rinishini keltiramiz:



Segmentlash va PoS teglash masalasi. Korpusni yaratish ko‘plab mashaqqat va vaqtni talab qiluvchi jarayondir. Avtomatik dasturlarning yordamisiz korpusni tuzish amrimahol. Ingliz tili

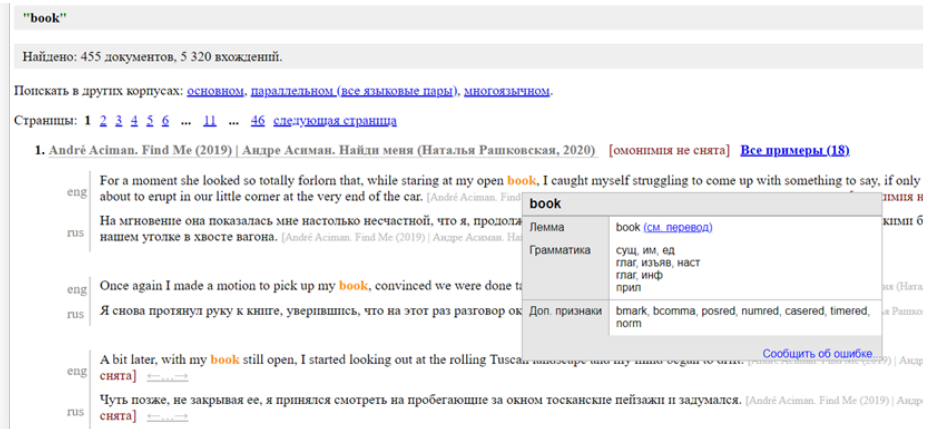
so‘zlarini PoS teglovchi dasturlarning bir necha turi allaqachon yaratilgan bo‘lsa, o‘zbek tili xususida bunday deb ayta olmaymiz. Agar o‘zbek tili uchun ham PoS tegger yaratilsa o‘zbek tili so‘zlari quyidagicha PoS teglanadi va segmentlarga ajratiladi:

```
<TEXT>
<TEXT HEAD>
        <FIELD>Hayvonot</FIELD><STYLE>Ilmiy</STYLE>
        <PERIOD>2022</PERIOD><U_TITLE>Sut emizuvchi hayvonlar</U_TITLE>
</TEXT HEAD>
<TEXT_BODY>
        <p id = “1”>
      <a id = “1” no = “1”>
        <s id = “1”>
           <U_TITLE><w pos = “t”>Hayvonlar</w>
                <w pos = “NOUN”>Arslon</w>
                <w pos = “CON”>va</w>
                <w pos = “NOUN”>sirtlonlar</w>
                <w pos = “NOUN”>sut emizuvchilar</w>
                <w pos = “VERB”>hisoblanadi.</w></U_TITLE></s></a></p>
         <p id = “2”>
```

Korpusdagi so‘zlarni morfologik teglashda ko‘pgina tillar uchun majburiy bo‘lgan 12 kategoriya mavjud [Zeroual, Lakhouaja, 2022, 61]:

Verbs	Fe’llar	VERB
Nouns	Otlar	NOUN
Pronouns	Olmoshlar	PRON
Adjectives	Sifatlar	ADJ
Adverbs	Ravishlar	ADV
Numerals	Sonlar	NUM
Determiners (determiners and articles)	Aniqlovchi (mas., artikllar)	DET
Adpositions (prepositions and postpositions)	Qo‘shimchalar (old va keyingi)	ADP
Particles	Qismlar	PRT
Conjunctions	Bog‘lovchilar	CON
Other (typos, abbreviations…)	Boshqa (xatolar, qisqartmalar)	X
Punctuation marks	Tinish belgilari	SENT

Bundan tashqari parallel korpusga lemmatizatsiya va tokenizatsiya funksiyasini qo‘shsak, korpusning foydalilik darajasi yanada ortadi. Quyida rus tili milliy korpusidan namuna keltiramiz:



Yuqoridagi ilovada rus tili milliy korpusi tarkibidagi ingliz-rus parallel korpusidan "book" so'zi qidirilmoqda. Avvalo, "book" so'zi qatnashgan parallel matnlar aks ettirilgan. Ko'rib turganimizdek, so'zning lemmasi "book", grammatik jihatdan ot va fe'l so'z turkumi, birlikda ekanligi ko'rsatilgan. Bu nafaqat "book" so'zi uchun, balki qidiruv natijasida paydo bo'lgan matnning istalgan, u xoh rus tili, xoh ingliz tili so'zining ustiga bosish orqali ko'rinadigan ma'lumotlardir. Demak korpusdagi so'zlarning har biriga lingvistik izoh berilib teglangan, bundan tashqari matnlarga ekstralingvistik teglar biriktirilgan. Qidiruv jarayonida qidirilayotgan so'zning nechta hujjatda va nechta o'rinda kelayotganligi ham aks etgan. Bir so'z bilan aytganda, rus tili milliy korpusidagi parallel korpuslarni mukammal teglangan deb sanash mumkin. Nafaqat lingvistik, balki yuqori sifatli ekstralingvistik teglanganlik xususiyati ham rus tili milliy korpusi ustida qanchalik ter to'kilganligini ko'rsatadi.

Korpusni indekslash masalasi. Korpusdan qidirish jarayoni vaqt talab etuvchi jarayon bo'lishi mumkin. Ammo, bu jarayonni tezlashtirish imkoni mavjud. Buning uchun qilinadigan ish ushbu korpusdagi matnlarning indekslanishidir. Bunda teskari indekslash usulidan foydalaniladi. Matnlar qismlarga bo'lib indekslanadi. Shunda matnning biror qismidagi ma'lumot qidirilayotganda, butun matn emas uning ozgina qismigagina murojaat qilinadi va qidiruv natijasi soniyalar ichida tezda namoyon bo'ladi. Agar yangi qo'shilayotgan matnlar ustida to'g'ri XML kodlash amalga oshirilsa, butun korpusni qayta kodlashni amalga oshirishga zarurat bo'lmaydi. Demak, indekslash yaratilishi kutilayotgan parallel korpusning mukammal qidiruv tizimini tashkil etuvchi omillaridan biri sanalmish tezkorlik talabini bajarishni o'z zimmasiga olar ekan, tatqiqotchilar o'zbek-ingliz parallel korpusini yaratishda ushbu jihatga alohida e'tibor qaratishlari maqsadga muvofiqdir.

Xulosa

Korpus lingvistikasi jadallik bilan rivojlanib bormoqda. Bu esa o‘z navbatida yurtimizda ham ushbu soha taraqqiyoti uchun harakatlarni ildamlashtirish masalalarini hal etishni taqozo qilmoqda. Buning uchun tashlanadigan ilk va istiqbolli qadamlardan biri – bu o‘zbek hamda jahon tillari juftligidagi parallel korpuslarni yaratish. Xususan, o‘zbek-ingliz tillari parallel korpusining yaratilishi tillararo ma’lumotlarni qayta ishlash, shuningdek, o‘zbek-ingliz tilli leksikografiya, til tadqiqoti va tilni o‘qitish uchun muhim manba bo‘lib xizmat qilishi shubhasiz, albatta. Ammo, bunday korpusni qanday yaratish, undan qanday qilib samarali ravishda foydalanish, unda qanday ma’lumotlarni jamlash kabi vazifalar dolzarb masala bo‘lib qolmoqda. Ushbu maqola korpus sohasida jahon tajribasiga tayangan holda, o‘zbek-ingliz parallel korpusini yaratish bosqichari to‘g‘risida nazariy ma’lumotlar beradi. Umid qilamizki, ushbu ish yurtimizdagi korpus sohasiga qiziquvchilar uchun dasturulamal bo‘lib xizmat qiladi va soha rivojiga munosib hissasini qo‘shadi.

Adabiyotlar:

Chanjun Park, Seolhwa Lee, Hyeonseok Moon Sugyeong Eo, Jaehyeong Seo and Heuiseok Lim. How should human translation coexist with NMT? Efficient tool for building high quality parallel corpus. 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia

John Hutchins. Machine translation and human translation: in competition or in complementation. International Journal of Translation, 13(1-2):5–20, 2001.

Jorge Leiva Rojo. Aspects of human translation: the current situation and an emerging trend. Hermeneus: Revista de la Facultad de Traducción e Interpretación de Soria, (20):257–294, 2018.

Q.F.Wen, L.F.Wang and M.C.Liang: Spoken and Written English Corpus of Chinese Learners[M], Beijing: Foreign Language Teaching and Research Press, 2005.(In Chinese)

CHANG Baobao. Chinese-English Parallel Corpus Construction and its Application. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo.

Gale, W. A., & Church, K. W. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(3), 75-102.

Karimov Rustam. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Dissertasiya. Buxoro – 2022

Imad Zeroual, Abdelhak Lakhouaja. MulTed: a multilingual aligned and tagged parallel corpus. Applied Computing and Informatics Vol. 18 No. 1/2, 2022 pp. 61-73

Alkahtani, Saad. Building and verifying parallel corpora between Arabic and English. DOCTOR OF PHILOSOPHY. Bangor University,

2015.
Leech, G.N. 2010. Corpus linguistics. In K. Malmkjaer (ed.), The Routledge Linguistics Encyclopedia, Third Edition. London: Routledge, 103-113.
Svartvik, J. 1992. Corpus linguistics comes of age. In J. Svartvik (ed.), Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter, 7-13.
Johansson, S. & Hofland, K. 1994. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie and P. Schneider (eds), Creating and Using English Language Corpora. Amsterdam: Rodopi, 25-37.
Johansson, S. 2007. Seeing through Multilingual Corpora. Amsterdam: Benjamins.

Steps of creating a tagged parallel corpus of the uzbek-english languages

Botir Elov[1]
Ma'rufjon Amirqulov [2]

Abstract:

The article talks about corpus linguistics, which is one of the main directions of computer linguistics, monolingual corpora and parallel corpora, and also about the stages of creating a parallel corpus of Uzbek-English languages based on world experience in the field of parallel corpora is maintained. In addition, information is provided about priority tasks such as establishing the programming and linguistic principles of the parallel corpus of Uzbek-English languages, linguistic and extralinguistic tagging of selected units, and developing an algorithm for creating a parallel corpus. Considerations are given on how to select data for the parallel corpus, the requirements for the data, and what opportunities the creation of the Uzbek-English parallel corpus provides to researchers and users.

[1]*Elov Botir Boltayevich* – doctor of philosophy in technical sciences (PhD), associate professor. Head of the Department of Computer Linguistics and Digital Technologies of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.
E-mail: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648

[2]*Amirqulov Ma'rufjon Aliqul o'g'li* – 2nd year master's degree in Computer Linguistics at Alisher Navoi Tashkent State University of Uzbek Language and Literature.
E-mail: amirkulovmaruf@navoiy-uni.uz
ORCID: 0000-0002-4025-8466
For reference: Elov B., Amirqulov M.. 2022. "Steps of creating a tagged parallel corpus of the uzbek-english languages". *Uzbekistan: language and culture. Applied philology*. 2 (5): 97-109.

In this process, linguistic and methodological problems, such as material selection, as well as programmatic difficulties in creating a parallel corpus are reflected.

Key words: *Corpus, tagging, alignment, parallel corpora, POS tagging, XML encoding.*

References:

Chanjun Park, Seolhwa Lee, Hyeonseok Moon Sugyeong Eo, Jaehyeong Seo and Heuiseok Lim. How should human translation coexist with NMT? Efficient tool for building high quality parallel corpus. 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia

John Hutchins. Machine translation and human translation: in competition or in complementation. International Journal of Translation, 13(1-2):5–20, 2001.

Jorge Leiva Rojo. Aspects of human translation: the current situation and an emerging trend. Hermeneus: Revista de la Facultad de Traducción e Interpretación de Soria, (20):257–294, 2018.

Q.F.Wen, L.F.Wang and M.C.Liang: Spoken and Written English Corpus of Chinese Learners[M], Beijing: Foreign Language Teaching and Research Press, 2005.(In Chinese)

CHANG Baobao. Chinese-English Parallel Corpus Construction and its Application. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo.

Gale, W. A., & Church, K. W. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(3), 75-102.

Karimov Rustam. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Dissertasiya. Buxoro – 2022

Imad Zeroual, Abdelhak Lakhouaja. MulTed: a multilingual aligned and tagged parallel corpus. Applied Computing and Informatics Vol. 18 No. 1/2, 2022 pp. 61-73

Alkahtani, Saad. Building and verifying parallel corpora between Arabic and English. DOCTOR OF PHILOSOPHY. Bangor University, 2015.

Leech, G.N. 2010. Corpus linguistics. In K. Malmkjaer (ed.), The Routledge Linguistics Encyclopedia, Third Edition. London: Routledge, 103-113.

Svartvik, J. 1992. Corpus linguistics comes of age. In J. Svartvik (ed.), Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter, 7-13.

Johansson, S. & Hofland, K. 1994. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie and P. Schneider (eds), Creating and Using English Language Corpora. Amsterdam: Rodopi, 25-37.

Johansson, S. 2007. Seeing through Multilingual Corpora. Amsterdam: Benjamins.