KORPUS LINGVISTIKASI

O‘ZBEK-INGLIZ TILLARI PARALLEL KORPUSIGA QO‘YILADIGAN LINGVISTIK VA EKSTRALINGVISTIK TALABLAR

Elov Botir Boltayevich
ToshDO‘TAU Ijtimoiy-gumanitar fanlar va axborot texnologiyalari fakulteti dekani, t.f.f.d. (PhD), dotsent,
Amirqulov Ma’rufjon Aliqul o‘g‘li
ToshDO‘TAU Kompyuter lingvistikasi mutaxassisligi 2-kurs magistranti

Annotatsiya: Maqolada korpus tilshunosligida muhim rol o‘ynovchi parallel korpuslar, xususan, o‘zbek korpusshunosligida yaratilishi zaruriyatga aylangan o‘zbek-ingliz parallel korpusiga qo‘yiladigan talablar xususida so‘z yuritiladi. Albatta korpusga, ayniqsa, parallel korpuslarga qo‘yiladigan talablar kundan-kunga ortib bormoqda. Quyida o‘zbek-ingliz parallel korpusiga qo‘yiladigan lingvistik va ekstralingvistik talablarga bafurja to‘xtalib o‘tilib, jahon korpusshunosligi tajribasiga tayangan holda ushbu talablarga qanday javob berilishi, bu jarayonlarni qanday amalga oshirilishi zarurligi haqidagi fikrlar beriladi.

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

Abstract: The article talks about parallel corpora, which play an important role in corpus linguistics, in particular, the requirements for the Uzbek-English parallel corpus, the creation of which has become a necessity in Uzbek corpus linguistics. Of course, the requirements for the corpus or corpora, especially for parallel corpora, are increasing day by day. Below, the linguistic and extra-linguistic requirements for the Uzbek-English parallel corpus will be discussed in detail, and based on the experience of world corpus studies, opinions will be given about how these requirements should be met and how these processes should be implemented.

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

Аннотация. В статье говорится о параллельных корпусах, играющих важную роль в корпусной лингвистике, в частности о требованиях к узбекско-английскому параллельному корпусу, создание которого стало необходимостью в узбекской корпусной лингвистике. Конечно, требования к корпусу или корпусам, особенно к параллельным корпусам, растут день ото дня. Ниже будут подробно рассмотрены лингвистические и экстралингвистические требования к узбекско-английскому параллельному корпусу, и на основе опыта изучения мировых корпусов будут даны мнения о том, как эти требования должны быть выполнены и как эти процессы должны быть реализованы.

Ключевые слова: *Корпус, тегирование, выравнивание, параллельные корпуса, POS-тегирование, XML-кодирование.*

Kompyuter lingvistikasining eng yuqori imkoniyatlariga ega yo‘nalishlaridan biri hisolanmish **korpus lingvistikasi** so‘nggi yillarda ko‘plab tatqiqotlar, izlanishlar markaziga aylanib ulgurdi. Ma’lumotlarning yozma shakldan elektron ko‘rinishga o‘tishi lisoniy birliklarni ham chetlab o‘tgani yo‘q, albatta. Bu esa til birliklari ustida bajariladigan vazifalarning bir qadar yengillashuviga olib keldi. Xususan, sarflanadigan vaqtning tejami bu afzalliklarning yorqin misollaridan biri bo‘lsa, birdan ortiq tillardagi lisoniy birliklarning o‘zaro elektron shaklda tenglashtirilishi tarjimashunoslik, qiyosiy lingvistika, lug‘atshunoslik hamda bir necha turli sohalardagi yangi yo‘nalishlarga keng yo‘l ochdi.

Korpuslar, ayniqsa, parallel korpuslar til va tarjima sohasining mashina tarjimasi, krosslingval ma’lumotlarni qayta ishlash, qiyosiy va tilga doir tadqiqotlar, tilni o‘rganish va o‘rgatish hamda bilingval leksikografiyada beqiyos o‘ringa ega hisoblanadi. O‘zining tilini yanada tanitish, uning xususiyatlarini barcha uchun qulay ko‘rinishga olib kelishda ham parallel korpuslar muhim ahamiyat kasb etadi. Shu jihatlarni hisobga olgan holda korpus lingvistikasiga bo‘lgan e’tibor so‘nggi o‘n yilliklarda, ayniqsa, ortib bormoqda.

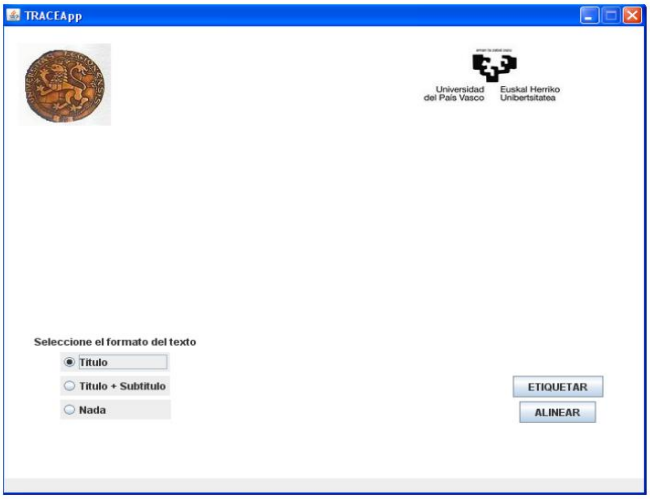
Yurtimizda kompyuter lingvistikasi sohasidagi izlanishlar, yutuqlar ko‘lami kundan-kunga ortib bormoqda va *“mamlakatimizda global ilmiy-texnik islohotlar amalga oshirilayotgan bugungi kunda kompyuter lingvistikasi doirasida avtomatik tarjima, o‘zbek tilini tushunish va qayta ishlash, uni boshqa tillar bilan chog‘ishtirishda sun'iy intellektning rolini belgilash, ikki va ko‘p tilli parallel korpuslar yaratishga kuchli ehtiyoj sezilmoqda.”*[1] Ushbu korpus va uning yaratilishi *“davlat tilining sofligini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining zamonaviy axborot texnologiyalari va kommunikasiyalariga faol integratsiyalashuvini ta’minlash”*[2]da o‘zining munosib hissasini qo‘shadi.

Parallel korpuslarning ilk elektron namunalari bundan 30 yil oldin yaratilishni boshlagan bo‘lsa, har bir til o‘zining lingvistik hamda dasturiy imkoniyatlaridan kelib chiqib parallel korpuslarni yaratish xususida fikrlar yuritishgan. Ba’zi tillar juftligidagi parallel korpuslar gap kesimida tenglashtirilgan bo‘lsa, ba’zilari hattoki so‘zlar kesimida ham tenglashtirilganiga guvoh bo‘lamiz. Ayrimlari lingvistik va ekstralingvistik teglashni maromiga yetkazgan bo‘lsa, qolganlari esa bu jihatlarga qisman e’tibor qaratishgan. Korpusning foydalilik koeffitsenti nafaqat uning qay darajada tenglashtirilgani bilan, balki uning qay darajada ekstralingvistik va lingvistik teglanganligiga ham bog‘liq.

Lingvistik teglash – matndagi ma’lumotlarning tilning grammatik, sintaktik, morfologik, semantik qoidalariga asosan izohlanishi.

Ekstralingvistik teglash – parallel korpusdagi matnlar haqida qo‘shimcha ma’lumotlar biriktirish, ularni izohlash. Masalan, matnning kimga tegishli ekanligi, qachon va qayerda yozilganligi, kim tomonidan tarjima qilinganligi, janri, so‘zlar soni va hkz.

Parallel korpuslarning ilk prototiplari faqatgina matnni abzaslar yoki gaplar kesimida tenglashtirishga qaratilgan bo‘lsa, hozirgi kunda yaratilayotgan korpuslarning imkoniyatlari bundan ancha keng ko‘lamda ekanligini kuzatishimiz mumkin. Quyida nisbatan oddiy darajada lingvistik va ekstralingvistik teglangan Aleuska(nemiz-ispan-bask) uch tilli parallel korpusi bilan tanishamiz. Naroa Zubillaga, Zurine Sanz va Ibon Uribarrilar[3] tomonidan yaratilgan korpus uchta tilni o‘z ichiga qamrab oladi. Korpusning maqsadi nemis tilidan bask tiliga qilingan bevosita tarjimalarni o‘rganish maqsadida yaratilgan. Bask tiliga ispan tilidan bilvosita qilingan tarjimalar ham o‘rin olgan. Korpusni yaratish jarayonida olimlar Trace-Aligner dasturini ham yaratishgan bo‘lib, ushbu dastur orqali matnlar abzas va gaplar kesimida tenglashtirilishi bilan birga, ular ekstralingvistik ma’lumotlar bilan ham boyitilganiga guvoh bo‘lishimiz mumkin:



1-rasm. Trace Aligner dasturi interfeysi.

Dastlab dastur orqali ma’lumotlarga ekstralingvistik izoh beriladi: avval matnlar avtomatik XML formatda annotatsiyalanadi, “header” qismga matn haqidagi metatavsiflar beriladi; keyingi jarayonda gaplar avtomatik ravishda tenglashtiriladi; so‘nggi bosqichda esa tenglashtirilgan avtomatik ma’lumotlarning to‘g‘rilik darajasi inson omili orqali tekshiriladi. Metatavsiflarni berish korpus uchun muhim omil deb sanalgan, ularning maqsadi korpus menejeri orqali so‘rovlar bajarilganda foydalanuvchilarga qulaylik yaratish bo‘lgan. Quyidagi rasmda Trace Aligner dasturi orqali tenglashtirilgan hamda ekstralingvistik teglangan ma’lumotlarni kuzatishimiz mumkin:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="spa">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Die Verwandlung</title>
<author>F. Kafka</author>
<translator>R. Iraola</translator>
<language>Deutsch</language>
<tradmode>Original</tradmode>
</titleStmt>
<publicationStmt>
<date>2011/08/08</date>
</publicationStmt>
<sourceDesc>
<p>Editorial_AGA:    Coleccion: </p>
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>
<p n="1"><s n="1">Die Verwandlung.</s></p>
<p n="2"><s n="2">Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte,  fand er sich in seinem Bett zu einem ungeheueren
</s><s n="3"> Er lag auf seinem panzerartig harten Rücken und sah,  wenn er den Kopf ein wenig hob,  seinen gewölbten,  braunen,
Versteifungen geteilten Bauch,  auf dessen Höhe sich die Bettdecke,  zum gänzlichen Niedergleiten bereit,  kaum noch erhalten konn
vielen,  im Vergleich zu seinem sonstigen Umfang klaglich dünnen Beine flimmerten ihm hilflos vor den Augen.</s></p>
<p n="3"><s n="5"><q n="1">"Was ist mit mir geschehen?",  dachte er.</q></s><s n="6"> Es war kein Traum.</s><s n="7"> Sein Zimmer,
```

2-rasm. Trace Aligner dasturi orqali XML formatda annotatsiyalangan matn.

Ko‘rib turganimizdek, matnlar abzas va gaplar kesimida tenglashtirilgan. Bundan tashqari matn haqidagi ekstralingvistik ma’lumotlar ham biriktirilgan: matn nomi, muallif, tarjimon, qaysi tilda ekanligi (original yoki tarjima), tarjima vaqti kabi.

Korpusning gaplar kesimida tenglashtirilganligi, unga ekstralingvistik metatavsiflar berilganligiga qaramay, lingvistik teglash jarayonining amalga oshirilmaganligi mukammal parallel korpusning shakllanishiga to‘sqin bo‘luvchi omil bo‘lgan. Zamonaviy korpusshunoslik rivojlanib borar ekan, korpuslarga, xususan, parallel korpuslariga qo‘yiladigan talablar ham ko‘payib boraveradi. Bunday talablarga mos keluvchi korpuslarni tuzish esa tadqiqotchilardan ko‘plab mashaqqatlarni, sabrni, ekstralingvistik, lingvistik hamda dasturiy bilimlarni talab qiladi. Bu esa mukammal parallel korpuslarning ko‘plab nusxalari yaratilishiga to‘siq bo‘lishi mumkin. Lekin bu kabi mashaqqatlar izlanuvchilarni harakatlardan to‘xtatib qo‘ygani yo‘q, albatta. Yillar o‘tgani sari parallel korpuslarga qo‘yiladigan talablarning ko‘lami ham ortib bordi. Yuqoridagi ta’kidlangan Aleuska parallel korpusidan farqli o‘laroq, 2004-yilda CHANG Baobao boshchiligida yaratilgan xitoy-ingliz parallel korpusi[4] so‘zlarning lingvistik izohlari sirasiga kiruvchi Pos teglash(so‘z turkumlari tegi) bilan boyitildi. Quyida ushbu teglanish qanday amalga oshirilganligini kuzatishimiz mumkin:

```
<TEXT>
<TEXT_HEAD>
    <MODE>书面语</MODE><FIELD>工商</FIELD><STYLE>新闻</STYLE>
    <PERIOD>当代</PERIOD><CH_TITLE>今年中国经济和社会发展八项任务</CH_TITLE>
</TEXT_HEAD>
<TEXT_BODY>
    <p id="1">
   <a id="1" no="1">
    <s id="1">
      <CH_TITLE><w pos="t">今年</w>
            <w pos="ns">中国</w>
            <w pos="n">经济</w>
            <w pos="c">和</w>
            <w pos="n">社会</w>
            <w pos="v">发展</w>
            <w pos="m">八</w>
            <w pos="q">项</w>
            <w pos="n">任务</w></CH_TITLE></s></a></p>
    <p id="2">
```

3-rasm. Xitoy-ingliz parallel korpusining ekstralingvistik va lingvistik teglanishi.

Ko‘rib turganimizdek, korpusda so‘zlarning qaysi so‘z turkumiga tegishliligini bildiruvchi teglar mavjud. Korpusni qurishda avtomatik dasturlardan foydalanilgan. Korpus tuzish jarayonida xitoy-ingliz tillaridagi matnlarni abzas va gaplar doirasida tenglashtiruvchi dastur, XML formatlovchi kodlash dasturi, xitoycha matnlarni segmentlarga bo‘lib Pos teglashni amalga oshiruvchi dasturlar yaratildi.

Alia Al-Sayed Ahmad, Bassam Hammo and Sane Yagilar boshchiligida yaratilgan ingliz-arab siyosiy parallel korpusini yanada mukammalroq yechimga ega korpuslar sirasiga kiritishimiz mumkin. Ushbu korpus[5] qirol Abdulloh II ning nutqlari, yozishmalari va bitta asarining arabcha hamda inglizcha talqinlaridan tashkil topgan. Ekstralingvistik teglanishning mukammal darajada amalga oshirilganligi korpusning ustunlik jihati bo‘lsa, yana biri uning yuqori darajada lingvistik izohlanishidadir. Korpus ustida metatavsif berish, matn segmentatsiyasi, tokenizatsiya, tenglashtirish, stemming va Pos teglash kabi jarayonlar olib borilgan. Quyida ushbu korpusdagi qirol Abdulloh II ning kitobiga biriktirilgan metatavsif bilan tanishamiz:

Title	**Our Last Best Chance: The Pursuit of Peace in a Time of Peril**
Publisher	Viking Press
Date of publication	2011
Place of publication	New York
No. of chapters	27
Chapter length (words)	2300-8194
English words	109491
Arabic title	فرصتنا الأخيرة: السعي نحو السلام في وقت الخطر
Publisher	Daralsaqi
Date of publication	2011
Place of publication	Beirut
Translator	Shukri Rahim
No. of chapters	27
Chapter length (words)	2300-8194
Arabic words	107634

1-jadval. Qirol Abdulloh II ning kitobiga biriktirilgan metatavsif.

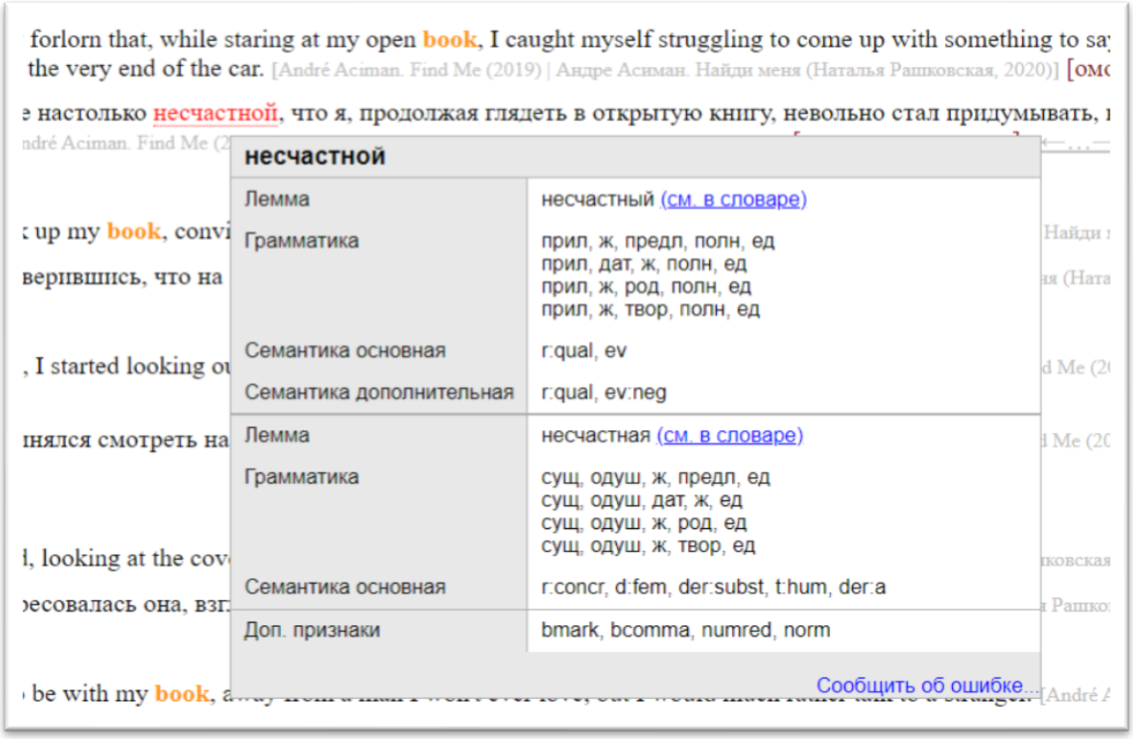
Nashr etuvchi, nashr vaqti, nashr joyi, boblar soni, boblardagi so‘zlar soni, inglizcha so‘zlar soni, arabcha nomi, tarjimon, arabcha so‘zlar soni kabilar ekstralingvistik izohlar sifatida tanlab olingan.

Lingvistik teglashda esa stemming, Pos teglash, so‘z shakllarini, so‘zlarning asosini biriktirishdan foydalanilgan. Bu esa korpus menejeri qidiruv tizimini mukammal ko‘rinishga olib kelgan. Chunki foydalanuvchi o‘zi xohlagan so‘zni stemlar, so‘z shakllar, asoslar, so‘z turkumlari va tokenlar doirasida qidirishi va natijada ko‘proq ma’lumotlarga ega bo‘lishi mumkin.

Biroq bundanda mukammalroq parallel korpuslarni ham uchratishimiz mumkin, albatta. Rus tili milliy korpusi hozirgi kundagi eng ulkan parallel

korpuslardan biri bo‘lib, nafaqat bu son jihatdan, balki sifatli (ekstra)lingvistik teglanganligi, foydalanuvchi uchun qulay interfeysga ega ekanligi bilan ham e’tiborga molik korpuslardan biridir. Parallel korpuslarining so‘zlar soni 168 millionga teng. Barcha parallel korpuslarda ham uchramaydigan sifatlarning biri shundaki, rus tili milliy korpusi va uning tarkibidagi parallel korpuslar mukammal lingvistik va ekstralingvistik izohlangan. Lingvistik teglash morfologik, sintaktik, semantik jihatdan amalga oshirilgan. Shu sabab korpus mukammal qidiruv tizimiga ega.

Guvohi bo‘lishimiz mumkinki, qidiruv tizimi kodlashdan umuman xabari bo‘lmagan foydalanuvchilar uchun ham juda qulay qilib ishlangan. Kiritilgan ma’lumot so‘z, so‘z shakli, leksema, jumla, semantik, grammatik birliklar yoki qo‘shimcha belgilar asosida qidirilishi mumkin. Quyida “несчастной” so‘ziga biriktirilgan teglarni ko‘ramiz:



4-rasm. Rus tili milliy korpusi tarkibidagi rus-ingliz parallel korpusidagi “несчастной” so‘ziga biriktirilgan lingvistik ma’lumotlar.

Xulosa o‘rnida aytish mumkinki, parallel korpuslarni yaratishda barcha uchun mos keluvchi, ma’lum qolipga tayanuvchi, bir xil lingvistik va ekstralingvistik teglanishni talab qiluvchi metod yoki yo‘l yo‘q. Biroq o‘zbek-ingliz tillari parallel korpusini yaratishda jahondagi eng mashhur va sifatli parallel korpuslar tajribasiga tayanish o‘zbek foydalanuvchilarini qoniqarli parallel korpus bilan ta’minlashda muhim rol o‘ynaydi. Ayniqsa, rus tili milliy korpusi tarkibidagi parallel korpuslar andozasiga tayanish ushbu maqsadlarga yetishda samaraliroq deb hisoblashni lozim ko‘ramiz.

Foydalanilgan adabiyotlar

1. R.Karimov. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. PhD ilmiy darajasini olish uchun yozilgan dissertatsiya. Buxoro-2022
2. O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoevning 2020 yil 20 oktyabrdagi «Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida»gi PF-6084-son farmoni // https://lex.uz/docs/5058351
3. Naroa Zubillaga, Zuriñe Sanz & Ibon Uribarri. 2015. Building a trilingual parallel corpus to analyse literary translations from german into basque. In Claudio Fantinuoli & Federico Zanettin (eds.), New directions in corpus-based translation studies, 71–93. Berlin: Language Science Press
4. CHANG Baobao. Chinese-English Parallel Corpus Construction and its Application. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo
5. Alia Al-Sayed Ahmad, Bassam Hammo and Sane Yagi. ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES. Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.
6. B.Elov, M.Amirqulov. O‘zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari. / O‘zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 90-102.
7. B.Elov, Sh.Hamroyeva, O.Abdullayeva, M.Uzoqova. O‘zbek tilida pos tegging masalasi: muammo va takliflar. O`zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 45-63.
8. O.Abdullayeva. Programs used to create the language corpus and their principles. ACADEMICIA: An International Multidisciplinary Research Journal. Vol. 10, Issue 6, June 2020. – B. 1778-1783.
9. https://ruscorpora.ru/