TIL KORPUSLARINING LINGVISTIK AHAMIYATI

Aybibi Iskandarova

O‘zbekiston Milliy universiteti

Kompyuter lingvistikasi va amaliy tilshunoslik kafedrasi dotsenti, f.f.n.

Tel: (99) 877 42 53

E-mail: aybibiiskandarova1962@gmail.com

Annotatsiya: *Ushbu maqolada til birliklarini raqamli saqlash, ularni empirik tahlil qilishda til korpuslarining lingvistik ahamiyati haqida fikr bildiriladi. Til korpuslarining parallellik nuqtayi nazaridan ikki va ko‘p tilli kabi turlarga ajratilishi, bunday korpuslarda asliyat hamda tarjima matnlarini qiyosiy tahlil qilish, tarjima birliklarining ekvivalentlarini kuzatish, badiiy asarlar elektron lug‘atini yaratishdagi ahamiyati haqida fikr bildiriladi.*

Kalit so‘zlar*: til korpuslari, parallel korpuslar, kontentlar hajmi, lingvistik baza, tarjima birliklari lug‘ati.*

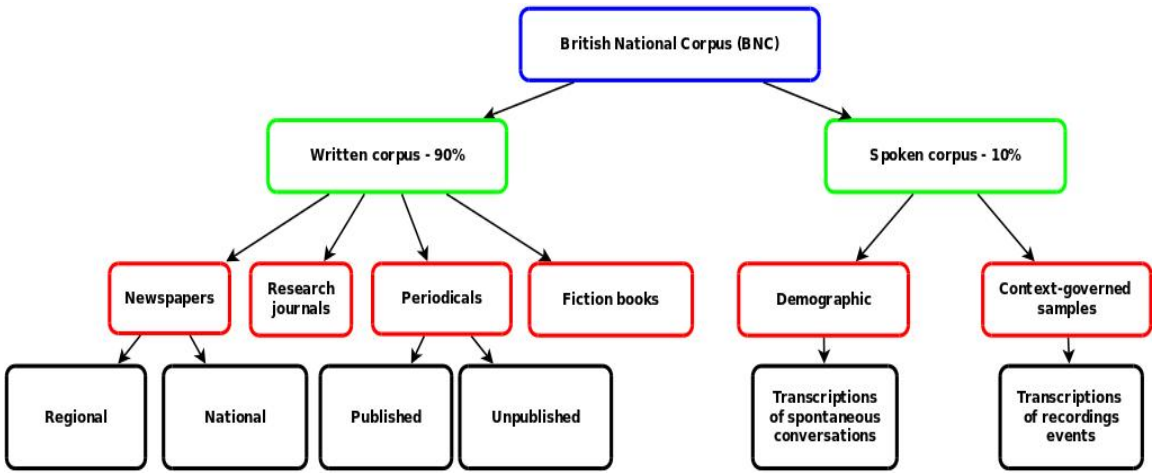
KIRISH

Korpus lingvistikasining paydo bo‘lishi va shakllanishi o‘tgan asrga borib taqaladi. XX asrda umumiy tilshunoslik fani tarkibida tabiiy tilning turli jihatlarini tadqiq qiluvchi: kompyuter lingvistikasi, psixolingvistika, sotsiolingvistika, lingvomadaniyashunoslik, lingvistik ekspertiza, neyrolingvistika kabi yangi yo‘nalishlar vujudga keldi. Til korpuslari til birliklarini raqamli saqlash va ularni empirik tahlil qilishning keng imkoniyatlarini yaratdi.

1990-yillarda dunyo tillarining kompyuter tahliligiga mo‘ljallangan 600ga yaqin: 1965-yilda–10ta; 1966-1970-yillarda–20ta,1971-1975-yillarda–30ta;1976-1980-yillarda–80ta; 1981-1985-yillarda–160 ta; 1986-1990-yillarda–320 ta til korpusi borligi qayd qilingan. Bugungi kunda minglab til korpuslari mavjud bo‘lib, ular maqsadi va til birliklarining hajmi bo‘yicha bir-biridan farqlanadi. Dunyodagi til korpuslari maqsadiga ko‘ra bir necha turlarga bo‘linadi:

- Milliy korpuslar (bitta til doirasida);
- Parallel korpuslar (ikki va undan ortiq tillar doirasida);
- Ixtisoslashgan korpuslar (tibbiyot, huquq, adabiyot);
- Og‘zaki nutq korpuslari;
- Tarixiy korpuslar.

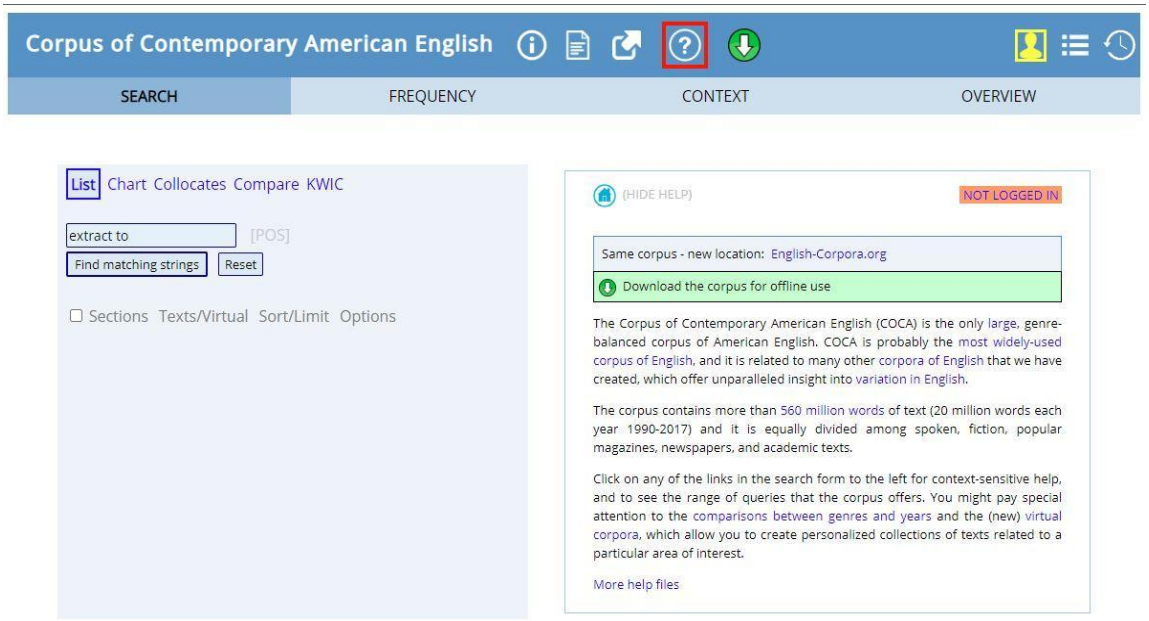
Britaniya milliy korpusi (BNC) va Amerika ingliz tili korpusi (COCA) ko‘p



sonli foydalanuvchilariga ega korpuslardan. BNC ning hajmi 100 million tokendan iborat, matnlar soni: 4000 ta, shulardan 90% yozma matn, 10% og‘zaki nutqdan iborat.

1. BNCning tarkibiy qismi

Amerika ingliz tili korpusi (COCA) ning hajmi 1 milliarddan ortiq tokendan iborat, u Britaniya Milliy Korpusidagi kontentlar hajmidan o‘n baravar katta. Bu ikkala korpus bitta til–ingliz tilining britan va amerika variantlari asosida shakllantirilgan bir tilli korpuslardir.



English	# words	language/dialect	time period	compare
iWeb: The Intelligent Web-based Corpus NEW!	14 billion	US/CA/UK/IE/AU/NZ	2017	Info (PDF)
News on the Web (NOW)	6.0 billion+	20 countries / Web	2010-yesterday	
Global Web-Based English (GloWbE)	1.9 billion	20 countries / Web	2012-13	
Wikipedia Corpus	1.9 billion	English	-2014	Info
Hansard Corpus	1.6 billion	British (parliament)	1803-2005	Info
Early English Books Online	755 million	British	1470s-1690s	
Corpus of Contemporary American English (COCA)	560 million	American	1990-2017	*****
Corpus of Historical American English (COHA)	400 million	American	1810-2009	**
Corpus of US Supreme Court Opinions	130 million	American (law)	1790s-present	
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	**
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	
CORE Corpus	50 million	Web registers	-2014	
Other languages				
Corpus del Español (see also...)	2.1 billion	Spanish	1200s-1900s	*
Corpus do Português (see also...)	1.1 billion	Portuguese	1300s-1900s	
N-grams				
Google Books: American English	155 billion	American	1500s-2000s	*
Google Books: British English	34 billion	British	1500s-2000s	
Google Books: Spanish	45 billion	Spanish	1500s-2000s	

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Korpusshunos olimlar parallel korpuslar (Parallel Corpora) haqida fikr

bildirib, ularni bir, ikki, ko‘p tilli kabi turlarga ajratadilar [1.20]. Parallel korpus – parallel tarjima matnlarining elektron analogi; ko‘plab “original matn va ularning bir/bir necha tarjimasi” bloklaridan iborat. Korpusdagi elektron matnlar original matnning o‘zi yoki uning bir qismi bo‘lishi mumkin [2]. Ikki, ko‘p tilli parallel korpuslarning xususiyati shundaki, ularda asliyat va tarjima matnlari mavjud bo‘ladi.

Zamonaviy korpus lingvistikasida parallel korpusning ikki ko‘rinishi qayd etiladi:

1) ko‘p tilli korpus (Comparable/Multilingual Corpora);

2) tarjima korpus (Translation Corpora)

Bunday xususiyatga ega korpusning struktur tarkibi ularning maqsadidan kelib chiqib turlicha bo‘lishi mumkin:

1) tarjimaga havola qiluvchi odatiy matn tarzida;
2) qiyoslash uchun qulay bo‘lgan “oynadagi matnlar” shaklida;
3) ma’lumotlar bazasi ko‘rinishida [3.81].

D.O.Dobrovolskiy RNCdagi parallel subkorpuslarning tuzilishi, matnlarning joylashuvi, tarjima birliklarining moslashtirilishiga ko‘ra quyidagi ko‘rinishlari borligini ta’kidlaydi:

1) bir yo‘nalishli (ingliz tilidan rus tiliga tarjima qilingan matn);

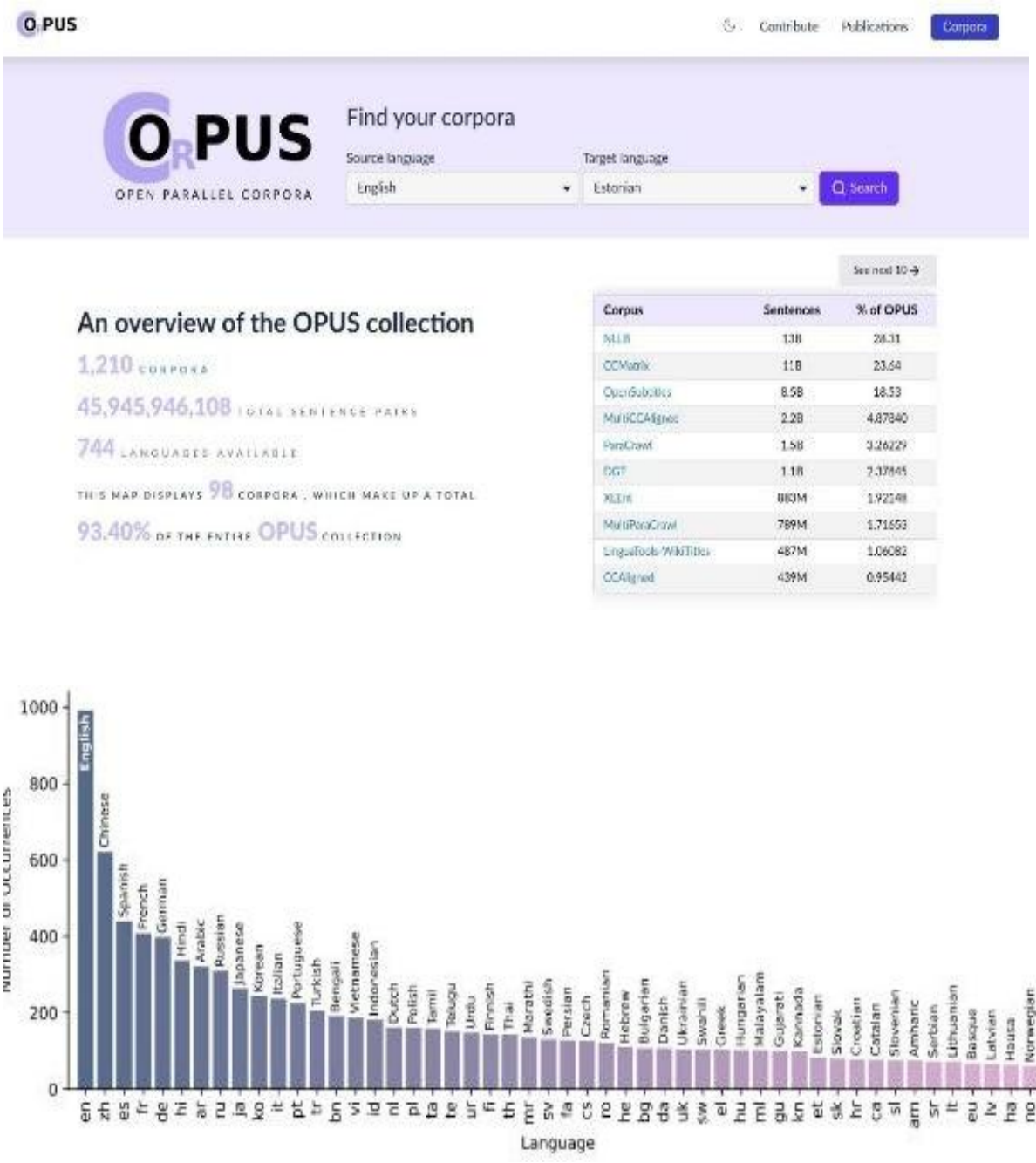
2) ikki yo‘nalishli (ingliz tilidan rus tiliga tarjima qilingan matn va rus tilidan ingliz tiliga);

3) ko‘p yo‘nalishli (ingliz tilidan rus, nemis, fransuz tillariga tarjima qilingan matn).

Ko‘p tilli “moslangan” parallel korpus namunasi sifatida Yevropa Ittifoqining Acquis Communautaire ma’lumotlar bazasini misol sifatida keltirishimiz mumkin. Har qanday parallel korpusning qimmati uning hajmi, tillararo matn jamlanmasi bilan belgilanadi.

Acquis Communautaire – Yevropa Ittifoqining huquqiy hujjatlaridan tashkil topgan ko‘p tilli parallel korpusi bo‘lib, OPUSning subkorpusidir. Uning ikki muhim jihati: korpusdan foydalanishning bepulligi, malti-eston, sloven-fin kabi

kam uchraydigan tilllar juftligining mavjudligi bilan baholanadi.



2. OPUS parallel korpusida qamrab olingan tillar soni

TAHLIL VA NATIJALAR

Parallel korpuslar moslangan (aligned) va moslanmagan (not aligned) shaklida bo‘lib, “moslangan” korpusda tarjima birliklari orasida bir-birini taqozo etuvchi aniq aloqa mavjud bo‘ladi. Moslangan korpusdan u yoki bu til birlklarining tarjima variantlari aniqlanadi. Bunday parallel korpuslar tarjimon uchun foydali bo‘lib, unda katta resurs–“tarjima xotira”si (Translation memory) mavjuddir. Badiiy

asarlardagi tarjima birliklari – ibora, ko‘p ma’noli so‘z, lakuna, realiya, ya’ni badiiy asarlardagi tarjima birliklari Translation memoryga yig‘iladi. Bu yig‘ilgan tarjima juftliklari boshqa asarlar tarjimasida qo‘l keladi, tarjimon uchun mos tarjima birligini tanlay olish imkoniyatini yaratadi va tarjima sifatini oshiradi. Moslangan korpus matnni tarjimasi bilan moslash, uning tarjimada qaysi birlikka tog‘ri kelishini tenglashdan iborat bo‘lib, bu jarayon avtomatik, unga qo‘shimcha ravishda qo‘lda bajariladi.

Biz tadqiqotimizda badiiy asarlarning asliyat va tarjima matnlari Smartcat CAT platformasi yordamida segmentlarga ajratildi. Mazkur dastur matnlarni segmentlarga bo‘lib, ularni yonma-yon joylashtiradi va natijada parallel matn juftligi (bitext) hosil qiladi.

Tahlil jarayonida ayrim hollarda o‘zbek lotin alifbosiga xos belgilar (o‘,q, g‘, h kabi) noto‘g‘ri aks etishi kuzatildi. Bu holat, asosan, matnning kodlash tizimi va apostrof belgilarining turlicha qo‘llanilishi bilan izohlandi.

Shuningdek, dastur segmentatsiyani punktuatsiya asosida amalga oshirgani sababli, qo‘shma gap tarkibidagi ayrim qismlar (masalan, vergulgacha bo‘lgan birliklar) alohida segmentlarga ajratildi. Shu bois segmentlar qo‘lda moslashtirildi.

Asliyat va tarjima matnlari o‘zaro muvofiqlashtirilgach, ular parallel korpusga joylashtirildi. Ushbu jarayon korpus lingvistikasida *alignment* termini bilan yuritiladi.

“Qiyomat” romanining asliyat va tarjima variantidagi gaplar bir-biriga CAT tool orqali dastlab quyidagi holda segmentlandi: Asliyatdagi bitta gap, tarjimada ikkita gap shaklida o‘girilgan.

И однако страх безрассуден, тем более уже знакомый, пережитый	***Бироқ қўрққанга қўшалоқ кўринади, деган гап бор.***
	Бунинг устига бир марта юрагингни олдирган сан

Keyin qo‘l mehnati bilan qaytadan ko‘rib chiqildi, gaplar bir-biriga moslangan holatda segmentlandi. Tarjimada ikkita gap asliyatga mos holatda bir

katakda joylashtirildi.

И однако страх безрассуден, тем более уже знакомый, пережитый	***Бироқ қўрққанга қўшалоқ кўринади, деган гап бор. Бунинг устига бир марта юрагингни олдиргансан***

	Очиқ дашт ёмон.
Это в открытой степи страшно, когда от преследующего вертолета некуда деться, когда он, настигая, неотступно гонится по пятам, оглушая свистом винтов и поражая автоматными очередями, когда в целом свете нет от вертолета спасения, когда нет такой щели, где можно было бы схоронить бедовую волчью голову, – ведь не расступится же земля, чтобы дать укрытие гонимым	***Унда тепангда қувалаётган вертолётдан ҳеч қаерга қочиб қутулолмайсан, у тинмай, орқангдан қолмай, изма-из таъқиб қилиб келаверади, вертолёт парракларининг касирғаси, автоматдан дўлдай ёғилган ўқларнинг чийиллаши юрагингни азоб ва қўрқинч билан тўлдиради, ёруғ дунёда вертолётдан омонлик йўқ, қочиб қутуладиган, шўрлик бошингни пана қиладиган тешик-туйнук ҳам йўқ — шундай пайтда ер ҳам ёрилмайди, қувиб келаётганлардан жонингни қутқариб, ерга кириб кетаман десанг.***

Bu misolda ham tarjimadagi ikkita gap asliyatga mos holatda parallel segmentlandi. Asliyat va tarjima matnlari to‘g‘ri segmentlanganidan keyin parallel korpusga joylanadi.

Это в открытой степи страшно, когда от преследующего вертолета некуда деться, когда он, настигая, неотступно гонится по пятам, оглушая свистом винтов и поражая автоматными очередями, когда в целом свете нет от вертолета спасения, когда нет такой щели, где можно было бы схоронить бедовую волчью голову, – ведь не расступится же земля, чтобы дать укрытие гонимым	***Очиқ дашт ёмон. Унда тепангда қувалаётган вертолётдан ҳеч қаерга қочиб қутулолмайсан, у тинмай, орқангдан қолмай, изма-из таъқиб қилиб келаверади, вертолёт парракларининг касирғаси, автоматдан дўлдай ёғилган ўқларнинг чийиллаши юрагингни азоб ва қўрқинч билан тўлдиради, ёруғ дунёда вертолётдан омонлик йўқ, қочиб қутуладиган, шўрлик бошингни пана қиладиган тешик-туйнук ҳам йўқ — шундай пайтда ер ҳам ёрилмайди, қувиб келаётганлардан жонингни қутқариб, ерга кириб кетаман десанг.***

XULOSA

Korpus lingvistikasining rivoji tilshunoslik sohalarining empirik bilish jarayonini yengillashtiradi: tadqiq etilayotgan til hodisalarini kuzatish, misol/dalil yig‘ish, taqqoslash, xulosa chiqarish uchun misollar massivining mavjudligi tahlil qilish bosqichini qulaylashtiradi.Til korpuslaridagi qulay interfeys va qidiruv tizimi tahlil maydonini kehgaytiradi. Shuningdek, korpusdagi tipik holatlar, misollarning o‘n, yuz, minglab uchrashi tadqiqotchiga xulosa chiqarishga yordam beradi; noodatiy holatlar esa yangi-yangi nazariy qarashlarni ham paydo qiladi. Aniqrog‘i, nutqda uchragan yangi birikma va struktura tilshunoslikda yangi farazlar paydo bo‘lishiga imkon beradi [5.]

Parallel korpuslar ikki til strukturasini tahlil qilishda, tarjima sohasida– original matn ekvivalentining boshqa tildagi adekvat variantini tanlashda; avtomatik tarjima va leksikografiya sohalari rivojida katta ahamiyatga egadir.

FOYDALANILGAN ADABIYOTLAR:

1. Захаров В.П., Богданова С. Ю. Корпусная лингвистика.– Иркутск: ИГЛУ, 2011.-Б.20.

2. Соснина Е.П. Параллельные корпусы в обучении языку и переводу. – Электрон ресурс:htt://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_education_translation/)

3. V. Zaxarov, B.Mengliyev, Sh.Xamroyeva. Korpus lingvistikasi: korpus tuzish va undan foydalanish. O‘quv qo‘llanma. T.2021.– 185 b.

4. Добровольский Д. Какие задачи решают параллельные корпусы. Электрон ресурс: https:// postnauka/ru/video/54851

5. Гёрн А. Параллельный корпус в системе университетских лингвистических курсов.– Электрон ресурс: http:// folk.uio.no/ atleg/nizhnij_gronn_oct2013.pdf

6. https://www.natcorp.ox.ac.uk/

7. https://www.english-corpora.org/coca/

8. https://opus.nlpl.eu/