IJTIMOIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNALI

ISSN 3060-5059

VOL.3 № 5 2026

JAHON LINGVISTIKASIDA KORPUSLASHTIRISH STANDARTI

Xolova Muyassar Abdulhakimovna
Termiz davlat universiteti, dotsent
Ro‘ziyeva Durdona
Termiz davlat universiteti, talaba

Annotatsiya
Ushbu maqolada mualliflik korpuslari ularga tegishli ma’lumotlar bazasini yig’ish doirasida fikrlar berilib, dunyo bo‘yicha turli, Broun, Flovn, Lob, Flob kabi korpuslar va ularning ishlash prinsiplari doirasida qarashlar kiritilgan.
Kalit so‘z: korpus lingvistikasi, Broun korpusi, Flovn, Lob, Flob korpuslari

СТАНДАРТЫ КОРПУСИРОВАНИЯ В МИРОВОЙ ЛИНГВИСТИКЕ

Холова Муяссар Абдулхакимовна
Термезский государственный университет, доцент
Рузиева Дурдона
Термезский государственный университет, студентка

Аннотация
В данной статье представлены идеи о сборе авторских корпусов и соответствующих им баз данных, а также рассматриваются различные корпуса со всего мира, такие как Brown, Flown, Lob, Flob, и принципы их работы.
Ключевые слова: корпусная лингвистика, корпус Brown, корпуса Flown, Lob, Flob.

CORPUS STANDARDIZATION IN WORLD LINGUISTICS

Xolova Muyassar Abdulhakimovna
Termez State University, Associate Professor
Ro‘ziyeva Durdona
Termez State University, Student

Abstract
This article presents ideas for collecting author-authored corpora and corresponding databases, and also examines various corpora from around the world, such as Brown, Flown, Lob, and Flob, and their operating principles.
Keywords: corpus linguistics, Brown corpus, Flown, Lob, and Flob corpora.

Korpus lingvistikasida “mualliflik korpusi” — bu lingvistik tahlil uchun belgilangan bitta muallifning barcha (yoki eng to‘liq to‘plami) matnlarining elektron to‘plami. Bunday manba idiolektni (individual uslubni) o‘rganish, mualliflik atributlari va tabiiy tilni qayta ishlash modellarini o‘qitish uchun muhimdir [1:178–182]. Mualliflar korpusi bitta yozuvchining matnlarini, ularning grammatik, semantik va uslubiy belgilarini, metama’lumotlarini va rivojlangan qidiruv tizimini o‘z ichiga oladi [2:1–6]. U muallif leksikonining “standarti” va muallif leksikografiyasi, chastota lug‘atlari va konkordanslarining asosi sifatida qaraladi [3:21–35].

Bibliyaga oid tadqiqotlar (Cruden lug‘ati), lug‘atlar (Johnson, Webster Dictionary), tillarni o‘qitish uchun (Thorndikning chastotali korpusi, 1921), Kvirk korpusi (Survey of English Usage) shular jumlasidandir. Kvirk korpusida 1 000 000 so‘z qo‘llanish holati mavjud bo‘lib, avvalo, har biri 17 satrli matndan iborat 4×6 dyum hajmli 1 000 000 kartochkadan tashkil topgan [4:331–334]. Mazkur korpus hozir London universitetida saqlanadi; o‘sha yerda undan foydalanish mumkin [5:114].

Inglizcha corpus linguistics atamasi ilk marta 1984 yilda qo‘llangan. Juda uzoq tarix hisoblanmasa-da, shu davr ichida korpus lingvistikasi zamonaviy tilshunoslikning peshqadam sohasi bo‘lib ulgurdi. Rossiyada ushbu atama korpus lingvistikasi mutaxassisi — mashhur xalqaro ingliz tili korpusi (International Corpus of English) asoschisi Sidni Grinbaum ma’ruzasidan keyin 1996 yilda kirib keldi. Ma’ruza tinglovchilaridan biri Integrum korpusida Sidni Grinbaumning korpus lingvistikasi bo‘yicha qilgan ma’ruzasiga tinglovchilar yog‘ilib kelganligi haqida qayd qilgan. Shu fakt rus korpus lingvistikasi tarixida ushbu tushunchaning ilk qo‘llanilishi bo‘ldi va keyinchalik atamaga aylandi. Albatta, korpus

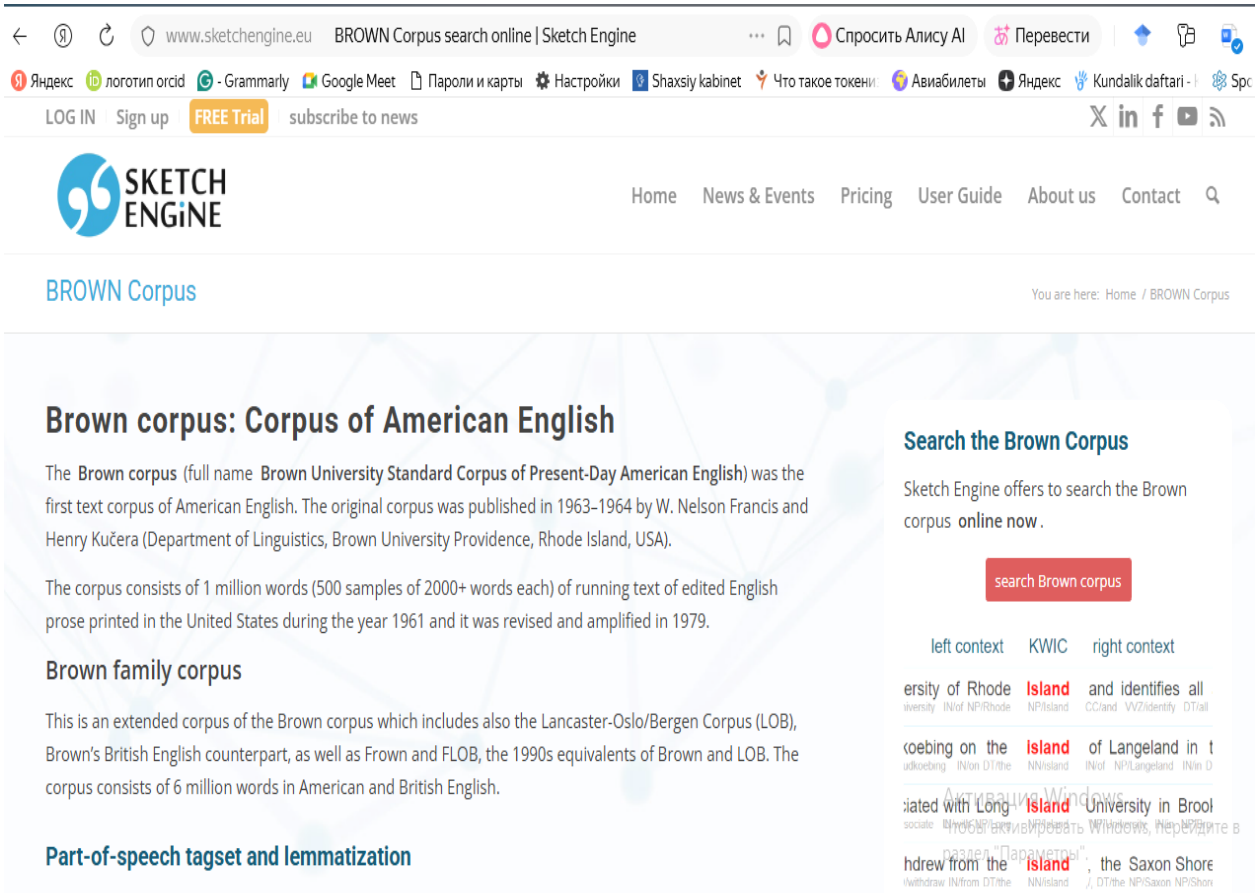
lingvistikasi yo‘q joydan paydo bo‘lgan emas. Tadqiqotning korpus metodi, korpus tuzish, undan foydalanishning ko‘p yillik tajribasi ham bunga asos bo‘ldi. Faqat kompyuter asrigacha bo‘lgan korpus noelektron shaklda bo‘lganligi, ma’lumotlarni korpusdan topishning avtomatlashtirilmaganligi bilan xarakterlanadi. Korpus lingvistikasida bu davr raqamli texnologiya (ing. pre-electronic) asrigacha bo‘lgan davr sanaladi. Mashhur tilshunos Panini tomonidan tuzilgan qadimgi hind grammatikasi shaklan noelektron, mohiyatan korpus metodiga asoslangan edi. Miloddan oldingi V–IV asrlarda bu korpus folklor shaklida og‘izdan og‘izga ko‘chib kelgan. Aslida, o‘sha paytdayoq o‘lik sanalgan sanskrit tilidagi vedalar matnidan iborat bo‘lgan [6:177 — https://www.myfilology.ru/177]. Kompyuter asrigacha bo‘lgan ko‘plab korpuslar ham turli diniy muqaddas kitoblar bilan bog‘liq.

Dunyo lingvistikasida XX asrning 60-yillarida korpus lingvistikasi doirasida tadqiqotlar olib borila boshlangan va bunda keng hajmli, katta massivli matnlar to‘plami vositasida ishonchli ma’lumotlar va faktlar olinishi mumkin degan nazariyalar Piatrovskiy R.G.ning miqdoriy belgilashga moslashtirilgan metodlar tilning yashirin qonuniyatlarini yuzaga chiqarishda ahamiyatli hisoblanib, zamonaviy lingvistikaning asosiy vositasiga aylanadi degan qarashlari orqali yuzaga keldi [7:216–224]. Biroq biz bu kabi izlanishlarga chuqurroq nazar solsak, korpusda maqsadli tadqiqotlar Blumfild va Bonjerlar tomonidan boshlangan. Undan keyin 1960-yilda turg‘un korpus hisoblangan Braun korpusi bu kabi ishlarning boshlanganiga yaqqol dalil bo‘la oladi, bu esa Frensis N. va Kucher G.larning korpus tuzishdagi dastlabki urinishi edi [8:34]. Ingliz tilidagi real matnlarga asoslangan COBUILD loyihasi asoschisi John Sinclairning lug‘atchilik borasidagi “portlash”i bo‘ldi va shundan so‘nggina korpus ustidagi ishlar jadallik bilan rivojlana boshladi [9:https://www.collinsdictionary.com/dictionary/engli]. Rus tilshunosligida A.B. Kutuzov, V.P. Zaxarov, E.V. Nedovshina, V. Plugnyan, V.V. Rikovlar korpusning turlari, xususiyati, tamoyillari, ilmiy va amaliy ahamiyati borasida tadqiqot olib borishgan (1.2.1-jadval) [10:101].

1-jadval

Korpus nomi	**Korpusning tuzilishi**	**Muallif / Tashkilot**	**Amaliy ahamiyati**
1964 — Brown Corpus	1 million so‘z, 500 ta matn (Amerika ingliz tili)	G. Kucher, V.N. Frensis (Brown University)	Zamonaviy korpus lingvistikasining boshlanishi, statistik til tadqiqotlariga asos bo‘lgan
1980-yillar — LOB Corpus	1 million so‘z, Britaniya ingliz tili matnlari	Lancaster–Oslo–Bergen guruhi	Amerika va Britaniya ingliz tilini qiyosiy o‘rganish imkonini berdi
1994 — British National Corpus (BNC)	100 million so‘z, yozma va og‘zaki matnlar	Oxford University Press, Longman va boshqalar	Leksikografiya, grammatik tadqiqotlar, til o‘qitish metodikasida keng qo‘llanadi
2008 — Corpus of Contemporary American English (COCA)	1 milliarddan ortiq so‘z, turli janrlar	Mark Davies (Brigham Young University)	Zamonaviy Amerika ingliz tilidagi o‘zgarishlarni kuzatish mumkin
2003 — Russian National Corpus	600+ million so‘z	Rossiya Fanlar Akademiyasi	Rus tilining tarixiy va zamonaviy tadqiqotlari
2011 — O‘zbek tilining milliy korpusi	Badiiy, ilmiy, publitsistik va og‘zaki matnlar	O‘zRFA Til va adabiyot instituti	O‘zbek tilini raqamli muhitda tadqiq qilish, avtomatik tahlil tizimlari yaratish

Braun korpusi (Brown Corpus). Amerika ingliz tili korpusi Braun korpusi (Brown universitetining zamonaviy Amerika ingliz tilining standart korpusi hisoblanadi). Bu Amerika ingliz tilining birinchi matn korpusi sifatida 1963–1964-yillarda V. Nelson Frensis va G. Kucher tomonidan nashr etilgan (Brown universitetining tilshunoslikka yo‘naltirilgan bo‘limi, Rod-Aylend, AQSh). To‘plam 1961-yilda Qo‘shma Shtatlarda chop etilgan ingliz nasrining tahrirlangan matnidan 1 million so‘zdan (har biri 2000 dan ortiq so‘zdan iborat 500 ta namuna) tashkil topgan bo‘lib, 1979-yilda qayta ko‘rib chiqilgan va to‘ldirilgan. Mazkur jamlanma Braun korpusining Britaniya inglizi analogi — Lancaster–Oslo/Bergen korpusi (LOB), shu bilan bir qatorda 1990-yillarda Braun va LOBning ekvivalentlari bo‘lgan FROWN va FLOBni ham o‘z ichiga oladi.



1-rasm. Braun korpusi interfeysi

FROWN — Freiburg-Brown Corpus of American English — Braun korpusining Amerika inglizi bo‘lib, 1961-yildan keyingi 30 yillik til o‘zgarishini taqqoslash asosida til evolyutsiyasiga ahamiyat berilgan. Hamda statistik nuqtayi nazardan talqin qilish imkonini taqdim etadi. FLOB esa Freiburg-LOB Corpus of British English bo‘lib, Britaniya ingliz tilining 1990-yillardagi yangilangan versiyasi hisoblanadi va bularni quyidagi jadvalda ko‘rish mumkin (1.2.2-jadval):

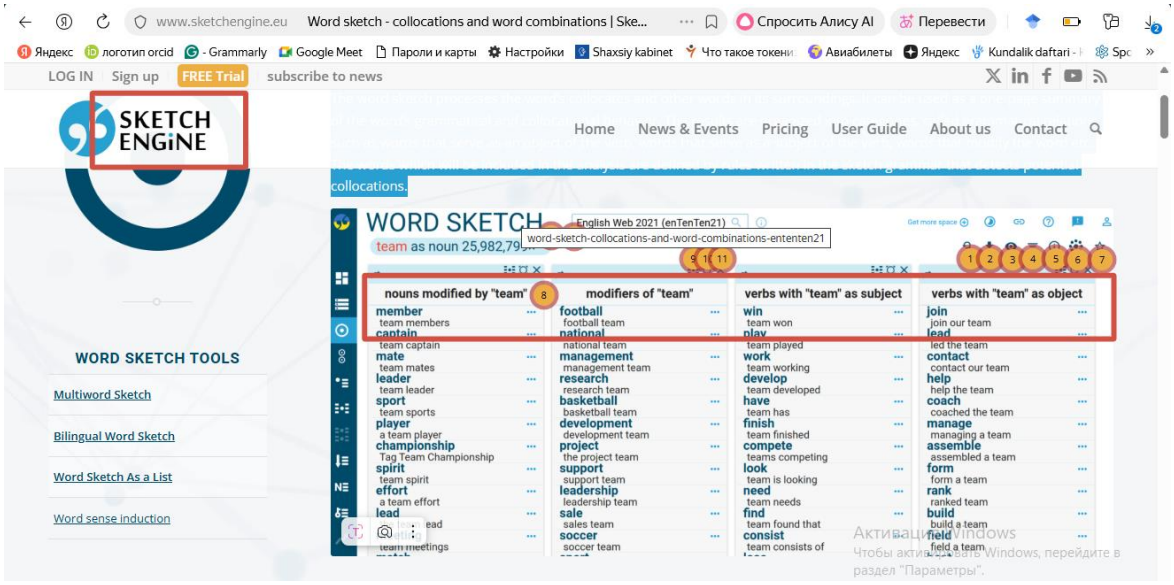
2-jadval.Brown korpusi tarkibi

Korpus qisqartmasi	**Yil**	**Til korpuslari**
Brown Corpus	1961	Amerika ingliz tili korpusi
LOB	1961	Britaniya ingliz tili korpusi
FROWN	1991	Amerika ingliz tilining takomillashgan korpusi
FLOB	1991	Britaniya ingliz tilining yangilangan korpusi

Korpus Amerika va Britaniya tillarida 6 million so‘zdan iborat. Unda teglar (annotatsiyalash) to‘plami va nutq qismlarini lemmatizatsiya qilish to‘plamlari mavjud. Braun korpusi — bu Penn Treebank (English Penn Treebank tagset [11:https://www.sketchengine.eu/penn-treebank-tagset/]) English Penn Treebank tagset ingliz tilidagi quyidagi teglar jamlanmasi bilan belgilangan nutqning bir qismi bo‘lib, unda nutq qismlari va grammatik kategoriyalar ko‘rsatib o‘tilgan.

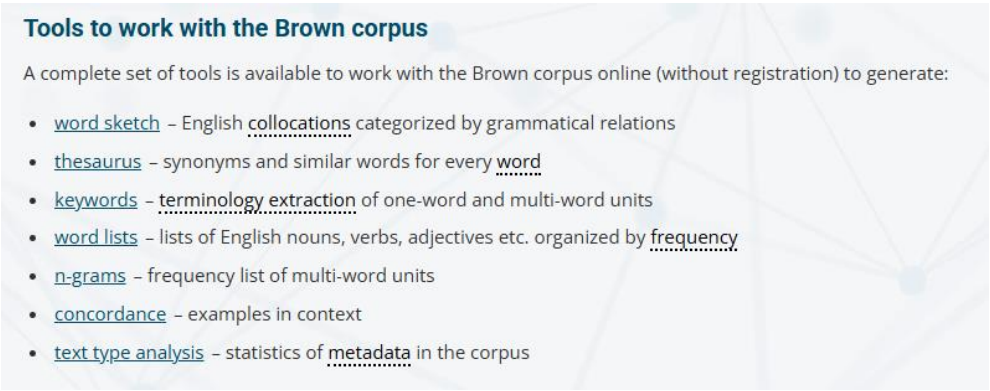
Korpus, shuningdek, CLAWS 7-versiyadagi teglar to‘plami hisoblangan POS teglariga ega. Korpus matnlari lemmatizatsiyani o‘z ichiga oladi, bunda korpusdagi har bir so‘z shakliga uning asosiy shakli (lemma) beriladi. Braun korpusi bilan ishlash vositalari. Braun korpusi bilan ishlash uchun to‘liq vositalar to‘plami onlayn (ro‘yxatdan o‘tmasdan) foydalanish uchun mavjud.

Word Sketch. Word Sketch’da ushbu so‘zning iboralari va uning atrofidagi boshqa so‘zlar qayta ishlanadi. U so‘zning grammatik va iboraviy xatti-harakatlarining bir sahifali xulosasi sifatida ishlatilishi mumkin. Natijalar grammatik munosabatlar deb ataladigan toifalarga bo‘linadi, masalan, fe’lning obyekti bo‘lib xizmat qiladigan so‘zlar, fe’lning subyekti bo‘lib xizmat qiladigan so‘zlar, so‘zni o‘zgartiradigan qismlar va hokazo. Bunda tahlilga kiritiladigan so‘zlar sketch grammatikasida yozilgan qoidalar bilan belgilanadi, bu esa potensial kollokatsiyalarni ochib beradi (1.2.2-rasm).



2-rasm. Braun korpusining Word Sketch oynasi

Yuqoridagi jarayonga sodda til bilan izoh berilganda: ogʻzaki namuna shakliy grammatik aloqalar boʻyicha tasniflangan inglizcha iboralar tezaurusi — har bir soʻz uchun sinonim va oʻxshash soʻzlar, kalit soʻzlar, bir boʻgʻinli va koʻp soʻzli til birliklarini terminologik topa olish, soʻzlar roʻyxati, ingliz tilidagi otlar, feʼllar, sifatlar va boshqa turkumlar roʻyxati, chastota boʻyicha tartiblangan n-grammalar, kengaygan birikmali tasodifiy bogʻlanishlarning chastotali roʻyxati, kontekstual misollar, matn turlarini tahlil qilish, tadqiqotdagi til birligining metamaʼlumotli statistikasi koʻrib oʻtilishi mumkin (1.2.3-rasm).



3-rasm. Braun korpusi vositalari

Biroq bu korpusning oʻziga xos jihati shundaki, undan roʻyxatdan oʻtgandan soʻng 1 oy bepul foydalanish mumkin, shundan soʻng esa pulli (37$) tarifga oʻtkaziladi va aksariyat bunday korpuslar (baʼzi ishlov beruvchi “app”larni hisobga olmaganda) muayyan toʻlov evaziga platformadan foydalanishga ruxsat beradi.Xulosa shuki, lingvistik korpuslar koʻp tarmoqli elektron korpuslarni oʻzida jamlagan multikorpus hisoblanib, u lingvistik va adabiy matnlarni tizimli ravishda toʻplash, raqamlashtirish hamda ilmiy asosda tahlil qilish imkonini beradi.

FOYDALANILGAN ADABIYOTLAR ROʻYXATI

1. Abjalova M., Gulomova N. Alisher Navoi Author’s Corpus: Value, Necessity and Significance // Proceedings of the 1st Pamir Transboundary Conference for Sustainable Societies (PAMIR). — 2024. — P. 178–182. — DOI: 10.5220/0012481400003792.
2. Abjalova M. et al. Alisher Navoi Author’s Corpus: Value, Necessity and Significance // Proceedings of the 1st Pamir Transboundary Conference for Sustainable Societies. — 2024. — DOI: 10.5220/0012481400003792.
3. Kalymon Y., Romanchuk O., Fedchyshyn N., Protsenko U., Yurko N. A corpus-based approach to author’s idiolect study: lexicological aspect // XLinguae. — 2022. — DOI: 10.18355/xl.2022.15.03.03.
4. Хамраева Ш.М. Особенности морфоанализа языковых корпусов // Страны. Языки. Культура. — 2020. — С. 331–334.

5. Кутузов А.Б. Корпусная лингвистика [Электронный ресурс]. — URL: Корпусная лингвистика — PDF
6. История корпусной лингвистики [Электронный ресурс]. — URL: История корпусной лингвистики
7. Пиотровский Р.Г. Лингвистическая статистика и автоматическая обработка текста. — М.: Высшая школа, 1981.
8. Френсис Н., Кучера Г. Вычислительный анализ современного американского варианта английского языка. — М., 1967.
9. Синклер Д. Предисловие к книге “Как использовать корпуса в преподавании иностранного языка” [Электронный ресурс]. — URL: Национальный корпус русского языка
10.Collins English Dictionary [Электронный ресурс]. — URL: Collins Dictionary (дата обращения: 15.02.2026).
11.Кутузов А.Б. Корпусная лингвистика [Электронный ресурс]. — URL: Корпусная лингвистика — PDF
12.Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров: учебно-методическое пособие. — Санкт-Петербург, 2006. — 26 с.
13.Рыков В.В. Курс лекций по корпусной лингвистике [Электронный ресурс]. — URL: Курс лекций по корпусной лингвистике
14.Плунгян В. Зачем мы делаем Национальный корпус русского языка? // Отечественные записки. — 2005. — № 2. — URL: Отечественные записки — статья Плунгяна
15.Penn Treebank Tagset [Электронный ресурс]. — URL: Penn Treebank Tagset