

**СЎЗ САНЪАТИ
ХАЛҚАРО ЖУРНАЛИ
4 ЖИЛД, 1 СОН**

**МЕЖДУНАРОДНЫЙ ЖУРНАЛ
ИСКУССТВО СЛОВА
ТОМ 4, НОМЕР 1**

**INTERNATIONAL JOURNAL
OF WORD ART
VOLUME 4, ISSUE 1**




14. Гафуров Бахтиёр АНАЛИЗ ЛИНГВОСТАТИСТИЧЕСКОЙ ХАРАКТЕРИСТИКИ АКЦЕНТНЫХ ФОНОВАРИАНТОВ ИМЕН СУЩЕСТВИТЕЛЬНЫХ РУССКОГО ЯЗЫКА.....	91
15. Ибаев Анвар КОМПАРАТИВ СИНТАКТИК ҚУРИЛМАЛАР ТИЛШУНОСЛИКНИНГ ЎРГАНИЛИШ ОБЪЕКТИ СИФАТИДА.....	97
16. Исмаилов Тургун ПРЕЗИДЕНТ ДИСКУРСИДА HOMELAND КОНЦЕПТИНИНГ МИЛЛИЙ ҚАДРИЯТЛАР ОРҚАЛИ ТАСАВВУР ЭТИЛИШИ (FREEDOM, LIBERTY ҚАДРИЯТИ МИСОЛИДА).....	103
17. Камолиддинова Вазира ОКСЮМОРОН ВА УНИНГ УСЛУБИЙ - СЕМАНТИК ХУСУСИЯТЛАРИНИНГ ҚЎЛЛАНИЛИШИ.....	111
18. Нигматова Лолахон ЎЗБЕК ТИЛИ МАХСУС КОРПУСЛАРИ – ТИЛ ВА МАДАНИЯТ МУШТАРАКЛИГИНИ ЎРГАНИШ ВОСИТАСИ.....	116
19. Сабирова Ёъзоза ЎЗБЕК ТИЛИНИНГ ИЗОҲЛИ ЛУҒАТЛАРИДА ЎЗЛАШМА ЮРИДИК ТЕРМИНЛАР ТАЛҚИНИ.....	121
20. Сулаймонова Нигора ЗАМАХШАРИЙ ГЎЗАЛ НУТҚ ВА САБР ОДОБИ ҲАҚИДА.....	130
21. Тўева Зулфия, Юлдашева Дилором БАЁН ВА УНИНГ ТУРЛАРИ ҲАҚИДА.....	137
22. Xolova Muyassar BOYSUN SHEVASI ONLAYN MA'LUMOTLAR BAZASI (QIDIRISH VA SARALASH IMKONIYATLARI ASOSIDA).....	146
23. Abdullayeva Oqila Xolmo'minovna O'ZBEK TILI KORPUSINI YARATISH BOSQICHLARI VA MUAMMOLI JHATLARI....	153
24. Abdunazarova Zarina Islomovna THE FORMATION OF SPEECH COMPETENCE.....	146
25. Nilufar Abduraxmonova O'ZBEK TILI KORPUSINI YARATISHDA LINGVISTIK ANNOTATSIYALASH TAMOIILLARI.....	164
26. Shamuradova Naima O'ZBEK VA INGLIZ XALQ MAQOLLARINING QIYOSIY O'RGANILISHI HAQIDA (ASOSIY QISMLARINI QO'LLASH).....	171
27. Яхёева Гулнора Бахтиёровна ОСОБЕННОСТИ РАЗВИТИЯ ТВОРЧЕСКОЙ АКТИВНОСТИ СТУДЕНТОВ В ПРОЦЕССЕ ОБУЧЕНИЯ ИНОСТРАННОМУ ЯЗЫКУ.....	175

СЎЗ САНЪАТИ ХАЛҚАРО ЖУРНАЛИ МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА INTERNATIONAL JOURNAL OF WORD ART

Abdullayeva Oqila Xolmo‘minovna,
Toshkent davlat o‘zbek tili va adabiyoti universiteti
tayanch doktoranti PhD

О‘ЗБЕК ТИЛИ KORPUSINI YARATISH BOSQICHLARI VA MUAMMOLI JIHATLARI

 <http://dx.doi.org/10.26739/2181-9297-2021-1-23>

ANNOTATSIYA

Maqolada o‘zbek tilining axborot matnlari korpusi tuzilishining o‘ziga xos nazariy va amaliy masalalari yoritib berildi. Korpusning tuzilishi, imkoniyatlari, dizayni, ish jarayoni bosqichma-bosqich tahlilga tortildi. Jumladan, o‘zbek tili korpusini qurish uchun texnik topshiriqlarni ishlab chiqish, ma’lumotlarni yig‘ish va ularni kompyuterlashtirish, korpusda matnlarni saqlash va so‘zlarni teglash jarayoni bayon qilindi. O‘zbek tili korpusi filologik jihatdan keng qamrovli va muammoli lingvistik masalalarni hal etish uchun mo‘ljallangan ma’lumotlar bazasidir.

Tayanch so‘zlar: korpus, teg, KWIC, konkordans, metama’lumot.

Абдуллаева Окила Холмуминовна,
Ташкентский государственный университет
узбекского языка и литературы PhD

ЭТАПЫ И ПРОБЛЕМЫ СОЗДАНИЯ УЗБЕКСКОГО ЯЗЫКОВОГО КОРПУСА

АННОТАЦИЯ

В статье рассматриваются конкретные теоретические и практические вопросы построения корпуса информационных текстов на узбекском языке. Пошагово анализировались структура, возможности, дизайн, рабочий процесс кейса. В частности, был описан процесс разработки Технического задания на построение корпуса узбекского языка, сбор и компьютеризация данных, хранение текстов в корпусе и разметка слов. Корпус узбекского языка представляет собой обширную филологическую базу данных, позволяющую решать проблемы.

Ключевые слова: корпус, teg, ключевое слово в контексте, конкорданс, метаданные.

Abdullaeva Okila Kholmo‘minovna,
PhD Researcher of Tashkent State University
of Uzbek Language and Literature

STAGES AND PROBLEMS OF CREATING THE UZBEK LANGUAGE CORPUS

ANNOTATION

The article deals with the specific theoretical and practical issues of the structure of the corpus of information texts of the Uzbek language. The structure, capabilities, design, work process of the case were analyzed step by step. In particular, the process of developing Terms of Reference for the construction of the Uzbek language corpus, data collection and computerization, storage of texts in the corpus and word tagging were described. The Uzbek language corpus is a comprehensive philological and problem-solving database.

Key words: corpus, tag, KWIC, concordance, metadata.

Kirish. Dunyo tilshunosligini kuzatib, korpus lingvistikasi va milliy korpuslarning yaratilishi oxirgi 20-30 yil ichida jadal o‘sganini xulosa qilish mumkin. Korpuslar uchun berilgan ta’riflar ichida Webster lug‘atida tilga oid ta’rifida “bilimlar yoki dalillar to‘plami; tilni tavsiflovchi tahlil uchun foydalanilgan so‘zlar to‘plami” [1] izohi berilgan, boshqa manbalarda korpus (ko‘pligi - corpora) yozma matnlar yoki yozib olingan og‘zaki nutqning transkripsiyasidan tuzilgan lingvistik ma’lumotlar to‘plami bo‘lib, asosiy maqsadi tilda mavjud farazlarni tasdiqlashdir deya ta’riflanadi [2, 85]. Korpus haqidagi ta’riflar va farazlar biroz bir - biridan farq qilsa-da, ammo, umumiy olganda, barcha fikrlar birlashib, korpus - til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo‘ysundirilgan matnlar majmui, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan matnlar jamlanmasi ta’rifini shakllantiradi.

Korpus termini qo‘llanilganda doim Milliy korpus birikmasi ishlatiladi. Milliy korpus – ma’lum bir tilning muayyan bir bosqichini ifodalaydigan barcha janrlar, uslublar, ijtimoiy hamda hududiy lahjalar va boshqalarda aks ettiruvchi elektron shakldagi matnlar to‘plami. Milliy korpus bo‘lishi uchun imkon qadar tilda mavjud yozma va og‘zaki matnlarning barcha turlarini o‘z ichiga olishi, korpusga kiritilib maxsus izohiga ega bo‘lishi kerak [3, 43]. Albatta bunda korpus hajmi boshqa korpus turlariga nisbatan oshib ketadi. Masalan, Ingliz tilining milliy korpusi (BNC) 100 milliondan oshiq, rus tilining milliy korpusi 200 millionga yaqin, turk tilining milliy korpusi 50 milliondan oshiq so‘z va so‘z shakllariga ega. Bugungi kunda dunyo tillarining juda ko‘pchiligi murakkab qidiruv tizimiga ega bo‘lgan, ilmiy va amaliy lingvistik tadqiqotlar uchun bir – biridan farqli tahlil funksiyalariga ega bo‘lgan milliy korpusiga ega. Shu qatori O‘zbek tili uchun ham milliy korpus yaratish tadqiqotlari tilshunoslar tomonidan boshlab yuborilgan. Bu borada o‘zbek olimlaridan A.Po‘latov, B.Mengliyev, N.Abdurahmonova, Sh.Hamroyeva, A.Eshmo‘minovlarning kompyuter lingvistikasi va korpus lingvistikasi sohalarida olib borayotgan ilmiy izlanishlarini e’tirof etish mumkin. Ammo ko‘pchilik ishlar natijasi nazariyalarni o‘rganish va tahlil qilish bilan cheklanib qolmoqda. Bir til doirasida milliy korpus yaratish murakkab va uzoq vaqtni oladigan jarayon ekanligi hech bir tilshunosga sir emas. Ammo kichik tadqiqotlar doirasida muayyan bir uslublar yoki janrlar bo‘yicha kichik korpus namunalari yaratish kerak. Lekin shu o‘rinda aytish kerakki, foydalanuvchi so‘zning kontekstlarda to‘g‘ri qo‘llanilishini tekshirishi yoki tabiiy til nutq birikmalarini izlashi, tilda tez-tez uchraydigan til qoliplarini aniqlashi va ulardan foydalanishi uchun kattaroq hajmdagi korpusga ehtiyoj sezadi.

Biz ingliz (BNC), rus (rucorpora) va turk tillarining milliy korpus (TNC)larini o‘rganib, ularda mavjud tajribalarni o‘zlashtirdik, ilmiy izlanishlar doirasida o‘zbek tilidagi internet axborot matnlari korpusini yaratish ustida tadqiqot olib boryapmiz. Bunda o‘zbek tilida mavjud xabarlarni tarqatuvchi saytlardan, jumladan kun.uz, daryo.uz, sof.uz, xabar.uz, darakchi.uz, gazeta.uz saytlaridan korpusga matn to‘planib, har bir so‘z va so‘z shakl maxsus izohlanadi va belgilar orqali teglanadi. Teglash jarayonida lemma va token masalasi ham hal qilinadi. Korpusdan foydalanish barcha qiziquvchilar uchun ochiq bo‘ladi, dasturdan tijoriy maqsad ko‘zlanmagan. Biz yaratayotgan korpus Maxsus korpus turiga to‘g‘ri kelsa-da, ammo kelajakda milliy korpus uchun fundament vazifasini bajarishiga ishonamiz.

Har bir tadqiqot ishida kuzatilgani kabi, til korpuslarini yaratish jarayoni ham ma’lum bir bosqichlari va muammoli jihatlari bilan murakkabdir. Avvalo o‘zbek tili korpusini yaratish uchun “Texnik topshiriqlar” dasturini tuzib oldik. Dasturda mutaxassis lingvist va dasturchi uchun

ko'rsatmalar berilgan. Mazkur ko'rsatmalar korpusni yaratishning bosqichma-bosqich amallari hisoblanadi. O'zbek tili korpusini ikki bosqichda amalga oshirdik. Birinchi bosqich dasturning Admin qismi, ya'ni korpusga matnlar kiritib, nutq qismlari tilshunoslik ma'lumotlari bilan izohlanadi yoki belgilanadi (tagging). Ikkinchi bosqich esa Foydalanuvchi qismi bo'lib, korpusda lingvistik tahlillar o'tkazadi. Masalan, so'zlar ro'yxatiga ega bo'lishi yoki eng ko'p qo'llanilgan so'zlar chastotasi haqida xulosa qilishi, real til matnlari turlariga ko'ra til qoliplarining ishlatilishi va nutq bo'laklaridan foydalanish vaziyatlari yuzasidan tahlil o'tkazish imkoniyatiga ega bo'ladi.

Korpus platformasi tayyor bo'lgach, matnlar yig'ila boshlanadi. Yozma matnlarni skanerlab, audio matnlarning transkripsiyasi olinsa, biz korpus uchun xabar saytlaridagi elektron matnlar tanlangani uchun qiyinchiliksiz ularni yuklab oldik. Olimlar korpusga to'plangan til namunalarini "xom" matnlar deb hisoblaydi (4, 31). Korpusning boshqa ma'lumotlar bazasidan yoki internetdagi elektron matnlaridan farqi til birliklariga maxsus lingvistik belgilar qo'yilishi, ya'ni teglanishi hisoblanadi. Korpusda nutq qismlarini belgilash yoki teglash nima? O'zbek tilida "yorliqlash" qanday amalga oshiriladi kabi savollar tug'ilishi tabiiy. Teglash korpusda kontekstdagi nutq qismlarining morfologik, semantik, sintaktik xususiyatlariga ko'ra standard belgilar bilan yorliqlanishidir. O'zbek tilida teg va teglamoq so'zining aynan muqobili yo'qligi sababli, bu termini belgilar qo'yilishi yoki belgi bilan yorliqlanishi tarzida qabul qildik. Biz tanlayotgan terminning qanchalik to'g'ri yoki noto'g'riligini xalq baholaydi. Nutq qismlarini korpuslarda teglash har doim murakkab jarayon sifatida baholangan. Chunki tillarning morfologik va semantik kategoriyalari umumiy va xususiy jihatlariga ko'ra o'z ichida yana turlarga ajratiladi. Biz korpusda kontekstlarda nutq qismlarini yorliqlash masalasini hal qilish uchun ingliz, rus va turk tillarining milliy korpuslaridagi teglash jarayonini o'rgandik. Shuningdek dunyo tilshunosligida yaratilgan CES (Corpus encoding standard) [5], TEI (Text encoding initiative) [6], CLAWS (Constituent likelihood automatic word-tagging system) [7] va Brill teglari [8] va teglash usullari bilan tanishdik. O'zbek tilida nutq qismlari uchun lotin grafikasi asosidagi standard qisqartmalarni olishga qaror qildik. Masalan, sifat so'z turkumi uchun <SIF> tegi, tinish belgilari uchun <TB> tegi yoki birinchi shaxs egalik uchun <1shE> teglari belgilandi. Masalan, kontekstda dorilarning so'zi lemma: dori <morf ot+Ko'p+Qark, sem AnOT>, keldim so'zi lemma: kel, <morf fe'l+O'tz+Xab+1sh, sem Jisf>, yozuvchining asari so'z birikmasi lemma: yozuvchi <morf ot+Bir+Qark+Yas, sem AOT>, lemma: asar <morf ot+Bir+3shE sem AnOt> sifatida teglanadi. Korpusning hozirgi imkoniyatlarida o'zbek tilidagi matnlar morfologik va semantik xususiyatlariga ko'ra annotatsiyalandi. Nutq birliklarini yorliqlashda yana bir muammoga duch kelindi, ya'ni o'zbek tilida lisoniy birliklar uchun teglar ishlab chiqilmagan va bir me'yorga solinmagan. Chunki teglar leksik yozuvlardan farq qiladi. Biz leksik qoidalarda mustahkamlangan birliklarning qisqartma nomlarini belgilashga harakat qildik. Nutq birliklarida omonimlik muammosi yuzaga kelmasligi uchun, kontekst o'rganilib, dasturda maxsus belgi orqali ajratildi. Barcha foydalanuvchilarga tushunarli bo'lishi, teglash jarayonida muammo yuzaga kelmasligi uchun o'zbek va ingliz tillaridagi standard qisqartmalarni yonma-yon kelishini ta'minladik.

Ba'zi olimlar aynan nutq qismlarini teglash jarayoni tabiiy tilni qayta ishlashning muhim qismi deb hisoblashadi. Chunki teglash jarayonida til birligining nafaqat nutqda bajarayotgan vazifasi va o'zi mansub kategoriyasiga ko'ra teglansa, boshqa tomondan nutq qismining kontekstdagi semantikasi va pragmatikasini ham tushunish va hisobga olish zarur hisoblanadi. Chunki bir o'rinda "ot" yorlig'i ostida kelgan nutq birligi boshqa o'rinda "fe'l" sifatida teglanishi mumkin. Hozir korpusda teglash va qidiruv jarayonini test sifatida bajarmoqdamiz. Nutq qismlarini teglayotganda til birligining kontekstdagi vazifasi, grammatik morfemalariga va kontekstual semantikasiga asosiy e'tiborni qaratdik.

Til korpuslarini yaratishda foydalanuvchi maxsus kompyuter dasturlari bilan tanishganimizda, ko'pchiligi standard usullarga ega ekanligining guvohi bo'ldik. Bu korpuslarda ma'lumotlarni aks ettirishning eng oddiy va oson usuli - so'zlar ro'yxati va chastotasini aniqlash funksiyasidir. Aslida ilk korpus namunalarida asosiy bosqich tabiiy til matnlarida uchrovchi so'zlarning ro'yxati va ularning nutqda takrorlanish sonini aniqlashdan iborat bo'lgan deb o'ylaymiz. Ammo korpus dasturlarida so'zlar ro'yxati ham bir xil emas. A.B.Kutuzov Korpus

lingvistikasi kursida soʻzlar roʻyxatining ikki turi mavjudligini yozgan. Yaʼni oddiy soʻzlar roʻyxati va konkordanslar roʻyxati [9, 29]. Oddiy soʻzlar roʻyxatida korpusda mavjud barcha soʻz va soʻz shakllari chastotasi koʻrinadi. Bu roʻyxat korpusda alfavit shaklida shakllantirilib berilishi mumkin, shuningdek korpus boyigani sari roʻyxat soni ham oshib boradi. Bu roʻyxatda soʻzlarning nechta matnda qoʻllanilgani va korpusda necha marta uchrashi statistikasi ham mavjud. Xuddi shu funktsiya korpusimizda mavjud. Ammo bu roʻyxat orqali soʻz shakllarining polisemiya va birmaʼnolilik xususiyatini koʻrishning iloji yoʻq, chunki kontekstsiz bu masalani yechib boʻlmaydi. Konkordanslar roʻyxati KWIC formatdagi soʻzlar roʻyxati deb ham nomlanadi. Yaʼni bu sodda roʻyxat emas, bunda soʻz kontekst bilan birgalikda beriladi. Bu formatda ham soʻz shakllarining chastotasi mavjud, faqat matnda soʻzning oʻng va chap tomondan boshqa birliklar bilan birikib kelish imkoniyatlarini ham tahlil qila olamiz. 1-rasmda “bilan” til birligining korpusda konkordans formatdagi tahlili koʻrsatilgan.

1-rasm. “bilan” soʻzining KWIC formatdagi tahlili

Snippet	Word	Context
... masala bo'yicha Braziliya tomoni	bilan	muntazam muloqotda bo'lib turibdi
... ushbu o'lim vaktsina sinovlari	bilan	bog'liq emasligini aytgandi
... Tibbiyot xodimlari kuni munosabati	bilan	koronavirus pandemiyasi davrida vafot
Koronavirus	bilan	kasallanganda o'pka zararlanganini tavsiflashda
... alveolalar tomirlardan kelgan suyuqlik	bilan	to'ladi
... alveolalar ham to'liq suyuqlik	bilan	to'lgan bo'lmaydi, deydi
... almashinuvi sezilarli darajada to'sqinlik	bilan	kechadi
Koronavirus	bilan	kasallanganda o'pkaning zararlanishi haqida
... mumkin (yoki pnevmoniya	bilan	emas, balki boshqa
... qilishi mumkin, shu	bilan	birga, o'pkada jiddiy
Shu	bilan	birga, kasallik belgilari
... Overseas Distribution LLC kompaniyasi	bilan	2019 yil 27 noyabrda
... dasturiy ta'minot ishlab chiqish	bilan	shu'ullanuvchi Zoom Video Communications
Zoom	bilan	birga Facebook
... pul mablag'larini naqdlashtirish	bilan	shu'ullanuvchi jinoyi guruh tuzgan
... shuningdek jinoyi yo'l	bilan	topilgan 200 million so'm
... o'z egalariga qaytarish vaji	bilan	moliya bo'limiga soxta xulosalarni
... qadar koronavirus yuqitirib olish	bilan	bog'liq 278 ta holat
... hamda koronavirusga chalingan bemorlar	bilan	muloqotda bo'lganligi sababli namuna
"Daryo" muxbiri	bilan	subhballashgan xususiy klinik mas'ullarining
... ota - onalariga murojaat	bilan	chiqdi
* Hammamiz koronavirus pandemiyasi	bilan	bog'liq ba'zi qiyinchiliklarni boshdan

Demak konkordanslar orqali soʻz birikmalari haqida xulosalar qilish mumkin. Mazkur funktsiya orqali oʻzbek tili matnlarida qaysi birliklar tez-tez yoki kamdan – kam qoʻllanilishini obyektiv baholashi, oʻzbeklar haqiqatdan ham tabiiy til birliklaridan qanday foydalanishini xulosa qilish mumkin. Masalan, muayyan miqdordagi matn namunalariga ega korpuslar orqali boshqa xalqlar ogʻzaki nutqida eng koʻp ishlatiladigan soʻzlar roʻyxati eʼlon qilingan. Shulardan biri Kembridj ingliz korpusining toʻrt yarim millionta ogʻzaki nutq namunalari asosida Shimoliy Amerikaliklar nutqidagi eng koʻp uchraydigan 500 talik soʻzlar roʻyxati eʼlon qilingan. Bu roʻyxatning eng boshida “men” (I) soʻzini koʻrishimiz mumkin [10]. Biz ham korpusimizda ikki xil formatda soʻzlar roʻyxati funksiyasini berdik. Umuman korpusda mavjud boshqa dasturlardan foydalanib ham, soʻzlar roʻyxati va konkordanslar, kollakatsiyalar roʻyxatini berish mumkin. Masalan, shunday tekin dasturlardan biri AntConc dasturlari boʻlib, mazkur dasturlar foydalanuvchi uchun qulayligi va universialligi bilan boshqa dasturlardan ajralib turadi. Bu dasturlarning afzalligi shundaki, muayyan til egalarining nutqiy faoliyatda nafaqat soʻz va iboralardan, maqollardan, shuningdek, soʻz yasash yoki soʻz shaklini hosil qilish uchun xizmat qiladigan lisoniy birliklardan foydalanish vaziyatlari va statistikasi boʻyicha qimmatli xulosalar olish mumkin boʻladi. Masalan, feʼlning oʻtgan zamon shaklini hosil qilishda – di va – gan qoʻshimchalaridan qaysi biri oʻzbeklar nutqida faol degan savol qoʻyilishi mumkin, sababi bu ikki qoʻshimcha doim ikki xil vaziyatni ifodalash uchun ishlatiladi deb qaraladi, yaʼni –di yaqin oʻtgan zamon, -gan uzoq oʻtgan zamon shakli uchun faoldir. Lekin nutqimizda ikkala qoʻshimchani aralash holatda ikkala vaziyatda ham qoʻllayveramiz. Demak oldingi faraz notoʻgʻri chiqishi mumkin. Aynan mana shu maqsadda korpusda statistika funksiyasini ham alohida berdik.

Mavjud tadqiqot ishlarini kuzatganimizda, korpus qurish va uning aniq balansi uchun ilmiy oʻlchov mavjud emasligini xulosa qildik (McEnery, Xino, Tono, 2006; McEnery, Hardie, 2012; Y.Aksan va boshqalar 2012). Til korpuslari oldingi mavjud korpuslar modellari orqali quriladi. Biz ham odatda rus tilining milliy korpusi modeli tizimidan foydalanishga harakat qildik. Ammo har bir tilning oʻziga xos ichki xususiyatlari mavjudligi korpusni qurish jarayonida murakkablik tugʻdirdi. Britaniya milliy korpusi, Turk tilining milliy korpusi va rus tilining milliy korpusi modellari

qiyosan o'rganilib, korpusning dasturiy ta'minoti qurildi. Korpusda ishlar bosqichma-bosqich amalga oshirildi. 1) o'zbek tilining barcha lingvistik xususiyatlari ma'lumot sifatida to'plandi; 2) to'plangan ma'lumotlar kompyuterlashtirildi va taxminiy to'g'ri sxemasi ishlab chiqildi; 3) mavjud elektron axborot matnlari yuklab olindi; 4) yuklab olingan matnlar kodlandi, ya'ni metama'lumotlar kirtildi; 5) matnlar kontekstidagi nutq qismlari annotatsiyalandi; 6) qidiruv tizimi ishlab chiqildi: bunda webga asoslangan barcha foydalanuvchilar uchun qulay interfeys yaratildi; 7) qo'shimcha ma'lumotlar qidiruv interfeysiga joylashtirildi: korpus, korpus imkoniyatlari va korpus mualliflari haqidagi ma'lumotlar; 8) korpusni e'lon qilish: korpusning eng oxirgi bosqichidamiz, ya'ni korpusni mahalliy sinovga chiqarish. Korpus versiyasi sinovdan muvaffaqiyatli o'tgandan so'ng xalqaro miqyosda foydalanilishi mumkin.

O'zbek tili axborot matnlari korpusi ma'lumotlarni yig'ish, saqlash, qayta ishlash va nazorat qilish uchun korpusni boshqarish tizimi ishlab chiqildi. Mutaxassislar korpus adminining ruxsati orqali olib boshqarish tizimidan ham foydalanish mumkin. Chunki korpusni qurish va ma'lumotlarni boshqarish ochiq va bepul dasturiy ta'minotga ega. Korpusni qurish va ma'lumotlarni yig'ish ishlarining dastlabki bosqichlarida turganligimiz uchun axborot matnlarining olinishida aniq bir davr belgilamadik. Bir so'z bilan aytganda, zamonaviy o'zbek tili korpusini qurishga qaratilgan loyihaning mahsulidir.

Istalgan foydalanuvchi korpusda qidiruv interfeysidan foydalanishida, turli lingvistik so'rovlarni amalga oshirishi uchun maxsus dasturlash tillarini bilishi shart bo'lmaydi, dasturda qulay va oson bo'lishi uchun maxsus qo'llanma ilova qilinadi, shuningdek ro'yxatlar hamda maxsus tugmalar bilan ta'minlanadi.

Foydalanuvchi qidiruv interfeysidan foydalanganda qidiruv natijalari oynada ko'rinadi. Oynaning o'ng tomonida mavjud natija statistikasi aks etadi, markazida barcha natijalarni ko'rish mumkin. Agar foydalanuvchi to'g'ridan – to'g'ri matn ustida amallar bajarmoqchi bo'lsa, asosiy matnning nomi ustiga bosiladi va to'liq matn olish mumkin. 2-rasmda covid so'zi qidiruvga berilgan. Hozirgi holatida korpus saytimiz test variantida bo'lganligi uchun saqlangan matnlar va ma'lumotlar ko'pchilikni tashkil etmaydi. Oynaning o'ng tomonida "aniqlandi" so'zi boshlanib, natija statistikasi mavjud. Bunda 7 ta matnda covid so'zi 22 marta uchragani ma'lum bo'ldi. Oynada so'rov aks etgan ma'lumotlarni ko'rish mumkin. Qidirilgan so'z rang bilan ajratilib berilgan. Har bir natijaning tepa qismida chiziqdan balandda matn nomi va undan yuqorida metama'lumotlar berilgan. Agar sichqoncha ko'rsatgichi bilan matn nomi ustida bosilsa, alohida oynada matnning to'liq matni namoyon bo'ladi. Undan yuqorida matnning metama'lumotlari ustiga bosilsa, matnning muallifi, matn e'lon qilingan sana va manbasini ko'rish mumkin.

2-rasm. Korpusda qidiruv natijasining ko'rinishi

Tadqiqotchi ma'lumotlarni saralashi, alifbo tartibida ustunlarda saqlash imkoniyatiga ega. Buning uchun natijalarni belgilab XLS (MS Excel) formatiga eksport qilish mumkin.

The screenshot shows a search interface for the word "COVID-19". The search results are displayed in a list format, with each entry showing the date, source, and a snippet of the text. The search results are sorted by relevance. The interface also includes a search bar at the top and a navigation menu on the right side.

Search results for "COVID-19":

- Aniqlandi: 7 ta matnda 22 ta so'z (Tizimda mavjud: 24 ta matn, 84 ta so'z)
- Dec 11, 2020 – www.gazeta.uz · oz · 2020 · 11 · 10 · covid-pneumo ... COVID-19 oqibatida o'pkaning zararlanishi qanchalik xavfli?...
- Bugungi kunda COVID - 19 gumon qilinganda yoki hatto aniq tashxis qo'yilgan biror kasallik hollarida ham ko'pincha kompyuter tomografiyasi (KT) amalga oshiriladi.
- Dec 11, 2020 – daryo.uz · 2020 · 09 · 05 · ozbekistonda-4-senta ... O'zbekistonda 4-sentabr kuni 278 kishida koronavirus aniqlandi. Kasall...
- O'zbekiston bo'yicha COVID - 19 tashxisi qo'yilganlarning 80, 8 foizi poytaxt hisobiga to'g'ri kelmoqda.
- Toshkent viloyatida 101 kishida COVID - 19 aniqlandi.
- Jizzaxda 21 - iyul kuni 29 kishida COVID - 19 qayd etilganidan so'ng biron marta kunlik kasallanish holatlari 8 taga yetmagandi.
- Samarqandda COVID - 19 qayd etilganlar soni 1745 nafarga yetdi.
- Andijon viloyatida 6 kishida COVID - 19 qayd etildi.
- Bu O'zbekiston bo'yicha COVID - 19 tashxisi qo'yilganlarning 1, 1 foiziga teng deganidir.
- Sentabr oyida Qashqadaryoda jami 21 kishida COVID - 19 aniqlandi.
- Farg'ona viloyatida 3 kishida COVID - 19 aniqlandi.
- Ma'lum bo'lishicha, marhum o'zida COVID - 19'ning barcha belgilari bo'lsa - da, test ham topshira olmagani, kasalxonaga ham yota olmagani.
- Sog'liqni saqlash vazirligi ma'lumotlari bo'yicha ayni paytda 2304 nafar bemor COVID - 19'dan davolanmoqda.
- Dec 11, 2020 – www.gazeta.uz · oz · 2020 · 11 · 10 · covid-19-usa ... AQSHda COVID-19 yuqtirganlar soni 10 milliondan oshdi...
- Jons Hopkins universiteti (AQSH) ma'lumotlariga ko'ra, AQSHda COVID - 19 bilan kasallanganlar soni 10, 1 million kishidan oshdi.
- Mamlakat bo'yicha shu kungacha qadar 238 mingdan ortiq inson COVID - 19 asoratlari sabab vafot etgan.
- « COVID tarqalishini to'xtatish uchun qila olishimiz mumkin bo'lgan eng samarali narsa bu niqob taqishdir.
- Dunyo bo'ylab so'nggi kun davomida COVID - 19 bilan kasallanganlar soni 495 mingga ko'paygan.
- Dec 11, 2020 – ... O'zbekistonda koronavirusga qarshi Xitoy vaktsinasining sinovi uchun ko'...
- Mazkur holatda organizmga virus va adenovirus emas, balki COVID - 19'ni keltirib chiqaruvchi koronavirus virusining oqsilli kiritiladi, deydi Abdullayev.
- Ko'ngillilar quyidagi asosiy talablarga mos kelishi lozim: bu 18 yosh oralig'idagi bo'lgan, jiddiy va surunkali kasalligi, allergiyasi yo'q, COVID - 19 bilan kasallanmagan va vaktsinatsiya vaqtida koronavirusga chalinmagan erak va ayollar.
- Dec 11, 2020 – daryo.uz · 2020 · 11 · 09 · koronavirusga-qarshi ... Koronavirusga qarshi vaktsina haqidagi yangiliklar sababli AQSH fond bi...
- AQSH fond birjatlari indeksi Pfizer kompaniyasi tomonidan yaratilgan COVID - 19 kasalligiga qarshi vaktsina haqidagi yangiliklardan so'ng rekord darajadagi kunlik o'sishni qayd etdi.
- Dec 11, 2020 – daryo.uz · 2020 · 11 · 09 · aqshning-pfizer-komp ... AQShning Pfizer kompaniyasi tomonidan yaratilgan koronavirusga qarshi ...

Korpus qurilishida, matnlar joylashuvida, albatta, xorijiy tillarda oldindan mavjud korpus amaliyotlari va tajribalaridan foydalanildi. Hali ish to'liq yakunlanmagan bo'lsa ham, lekin jarayon tadrijiy va bosqichma-bosqich amalga oshirildi.

Xulosa. Til korpuslarini yaratish modeli va korpuslar hal qiladigan lingvistik masalalar o'xshash bo'lsa-da, ammo korpuslarning ichki xususiyatlari, til qoliplari, lemma, token va teglash jarayoni kabi o'ziga xos muammoli jihatlari bilan farqlanib turadi. O'zbek tilining axborot matnlari korpusi keyingi 10 yillikda jadal rivojlantirilib, o'zbek tilining dunyo tillari qatorida munosib o'rin egallashida, tilda zamonaviy axborot texnologiyalaridan samarali foydalanishda alohida ahamiyat kasb etishiga, shuningdek o'zbek tilining lingvistik xususiyatlariga, boy xarakteriga qiziqqan har bir tadqiqotchi uchun muhim manba bo'lishiga ishonamiz.

Foydalanilgan adabiyotlar ro'yxati

1. "Corpus." Merriam-Webster.com Dictionary, Merriam-Webster, <https://www.merriam-webster.com/dictionary/corpus>.
2. Crystal, David. An Encyclopedic Dictionary of Language and Languages. Oxford, 1992.
3. Ю.А.Волоснова. Корпусная лингвистика: проблемы и перспективы. Лесной вестник 7/2006.
4. McEnery T., Xiao R., Tono Y. Corpus-based Language studies. Routledge, 2006.
5. CES <https://www.cs.vassar.edu/CES/>
6. TEI <https://tei-c.org/>
7. CLAWS <http://ucrel.lancs.ac.uk/claws/>
8. Brill tagger
https://web.archive.org/web/20090425061222/http://cosmion.net/jeroen/software/brill_pos/
9. Курузов А.Б. Корпусная лингвистика. 2015. http://tc.utmn.ru/files/corpus_5.pdf
Corpus frequency. <https://www.cambridge.org/gb/files/5913/9100/8829/Touchstone-2-Top-500.pdf>
11. Aksan Y., Aksan M., Koltuksuz A., et al. Construction of the Turkish National Corpus (TNC). <https://www.researchgate.net/publication/265914832>
12. <https://ruscorpora.ru/new/>
13. <https://www.english-corpora.org/bnc/>
14. <https://v3.tnc.org.tr/login>