ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 4 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 4 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Bosh muharrir: **Botir Elov**

Bosh muharrir o‘rinbosari: **Shahlo Hamroyeva**

Mas’ul kotib: **Oqila Abdullayeva**

Tahrir kengashi

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 19-27

CLARIN TIZIMIDAGI OG‘ZAKI KORPUSLAR XUSUSIDA

Nilufar Muradova[1]

Annotatsiya. Raqamli texnologiyalarning rivojlangani barcha sohalarda o‘z aksini topmoqda. Xususan, tilshunoslikda til korpuslarini yaratish, tabiiy tilga ishlov berish (NLP), mashina tarjimasi masalalari dolzarb. Albatta, bu tadqiqotlar tilimizning rivojlanishi va yashovchanligini oshirishga xizmat qiladi. Bunda, asosan, til korpuslari ahamiyatlidir. Dunyo tilshunosligida mukammal, kengaytirilgan qidiruv imkoniyatiga ega yirik korpus tizimlari ishlab chiqilgan. Bulardan biri Clarin tizimidir. Clarin – til ma’lumotlarini kashf qilish, o‘rganish, izoh qo‘shish, tahlil qilish, birlashtirish va lingvistik tadqiqot o‘tkazish imkonini beradi. Bu tizimga bir qancha til korpuslari ham kiritilgan. Ushbu maqolada Clarin tizimining maqsad, vazifa, imkoniyatlari; tizimdagi korpus, subkorpus, shuningdek, og‘zaki korpuslar tavsiflangan. Og‘zaki korpuslarning turlari, imkoniyatlari va qidiruv tizimi bayon etilgan. Xususan, chex tili og‘zaki korpusi haqida umumiy ma’lumot, korpusning ishlash tizimi, subkorpusdan farqi, qidiruv imkoniyatlari o‘rganilgan.

Kalit so‘zlar: *Clarin tizimi, korpus, og‘zaki korpus, lemmatizatsiya, qidiruv tizimi.*

Kirish

Korpus lingvistikasi jahon kompyuter lingvistikasining juda tez rivojlanib kelayotgan sohasi bo‘lib, bu borada ancha yutuqlarga erishilgan. Oliy ta’lim muassasalarida korpus lingvistikasi fan sifatida ham o‘qitiladi. Bu sohaning predmeti korpus yaratish nazariyasi va amaliyoti bo‘lsa, fan sifatida korpusning o‘ziga xosligi, dasturlash asoslari kabi jihatlari o‘qitiladi. Korpus lingvistikasi kompyuter lingvistikasining tarkibiy qismi bo‘lib, til korpusini yaratish, kompyuter texnologiyasi yordamida ulardan foydalanishning umumiy nazariyasi va amaliyoti bilan shug‘ullanadi [Захаров, 2005. 48]. Dunyo tilshunosligida korpus tuzish, foydalanish va uni ishlab chiqish tamoyillari rivojlangani sari nafaqat yangi tizimlar

[1]*Muradova Nilufar Baxdir qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti kompyuter lingvistikasi mutaxassisligi magistranti.
E-pochta: *muradovanilufar06@gmail.com*
ORCID: 0009000184741863

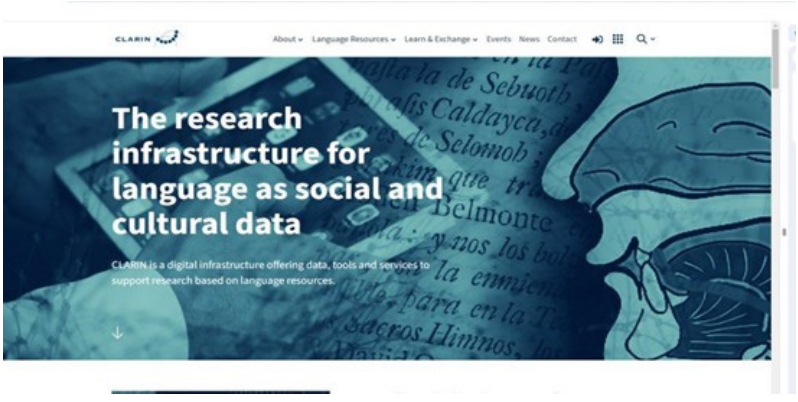
balki, mukammal korpuslar ham ishlab chiqilmoqda. Mana shunday tizimdan biri Clarin korpuslar jamlanmasidan iborat tuzilmadir. Clarin - til ma'lumotlarini kashf qilish, o'rganish, izoh qo'shish, tahlil qilish, birlashtirish va lingvistik tadqiqot o'tkazish imkonini beradi. Bu dastur 2019-yil fevral oyida tashkil qilingan (https://www.clarin.eu/). Clarin - bu til manbalariga asoslangan tadqiqotlarni qo'llab-quvvatlash uchun ma'lumotlar, vositalar va xizmatlarni taklif qiluvchi tizim.

Clarin tizimining maqsadi butun Yevropa va undan tashqaridagi barcha raqamli til manbalari va vositalariga gumanitar va ijtimoiy fanlar tadqiqotchilarini qo'llab-quvvatlash uchun yagona onlayn muhit yaratish. Clarinning vazifasi gumanitar va ijtimoiy fanlar bo'yicha tadqiqotlar uchun til ma'lumotlari va vositalarini almashish, ulardan foydalanish va barqarorligini qo'llab-quvvatlash uchun tizimlarni yaratish va saqlashdir.

Asosiy qism

"Og'zaki korpuslar" o'zbek tilida bo'lib, ingliz tilida "oral corpora" deb tarjima qilingan. "Og'zaki korpus" lingvistik tahlil uchun ishlatiladigan og'zaki til namunalari to'plamini anglatadi, asosan suhbatlar, intervyular, nutqlar yoki og'zaki muloqotning boshqa shakllarini yozib olish. Bu korpuslar tilning fonetika, sintaksis, semantika, nutq va sotsiolingvistika kabi turli jihatlarini o'rganish uchun qimmatli manbadir.

Clarindan foydalanish va unga til manbalari va vositalarini joylashtirish uchun tizimdan ro'yxatdan o'tish hamda a'zo bo'lish kerak. Hozirgi kunda 23 ta shahar va davlatlar ro'yxatdan o'tgan va a'zo bo'lgan. Bular: Avstriya, AQSH, Buyuk Britaniya, Belgiya, Bolgariya, Xorvatiya, Kipr, Chexiya, Daniya, Estoniya, Finlandiya, Germaniya, Gretsiya, Vengriya, Islandiya, Italiya, Latviya, Litva, Niderlandiya, Janubiy Afrika, Shvetsariya, Kolumbiya, Portugaliya. Clarin tizimining bir qancha markazlari mavjud. Markazning 3 ta turi mavjud: B markazlari, C markazlari va K markazlari. Ularning har biri turli xil tajriba va xizmatlarni taklif etadi. Maxsus Clarin markazlari (Clarin centre) 50 ga yaqinni tashkil etadi. Barcha tashkil etilgan Clarin markazlarining to'liq ro'yxati https://www.clarin.eu/content/overview-clarin-centres saytida mavjud. Clarin tizimi 6 bo'limdan iborat (1-rasm).



1-rasm. Clarin tizimi interfeysi

Birinchi bo‘limda (About) tizim haqida umumiy ma’lumot, Clarin tizimining markazlari, boshqaruv tizimi, a’zo bo‘lgan davlatlar va ishlash texnologiyasi haqidagi dastlabki ma’lumotlar aks etgan.

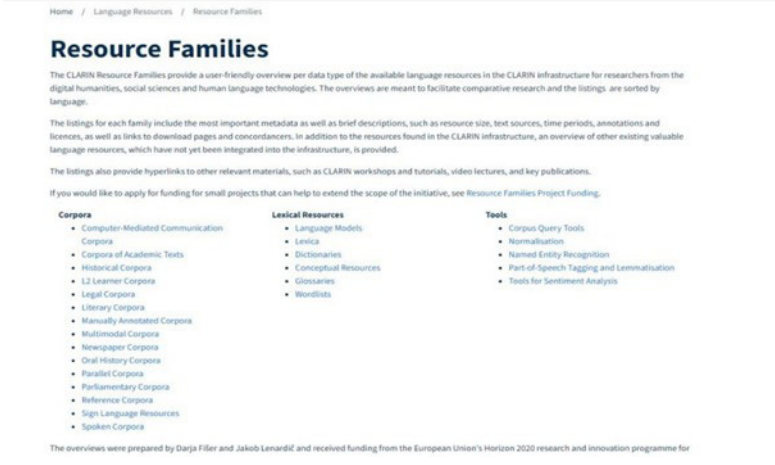
Ikkinchi bo‘limda (Language Resources) foydalanish uchun qulay til manbalari. Bu yerda siz Clarin tizimining til manbalariga osongina kirishingiz mumkin. Nutq va til ma’lumotlarining keng turlarini, shuningdek, ma’lumotlarni qayta ishlash uchun dasturiy vositalar va xizmatlardan foydalanishingiz mumkin. Yozma va og‘zaki korpuslar, shu jumladan, multimodal resurslari va ma’lumotlar bazasi mavjud. Undan tashqari Clarin til ma’lumotlarini izohlash, tahlil qilish yoki birlashtirish uchun turli xil vositalar va xizmatlarni taklif etadi. Taklif qilinayotgan narsalarni ko‘rib chiqish yoki ehtiyojlaringizga mos keladigan maxsus vositalarni tanlash mumkin. Shuningdek, tizimdagi mavjud til resurslarining ma’lumotlar turi bo‘yicha foydalanuvchilarga qulay ma’lumot beradi.

Uchinchi bo‘limda (Learn Exchange) amaliy tadqiqotlar, foydalanuvchilarni jalb qilish bo‘yicha mavjud tadbirlar, ishlab chiqilgan vositalar va tizimdan foydalanayotgan taniqli tadqiqotchilar bilan suhbatlar joylashgan.

To‘rtinchi bo‘limda (Events) tadbirlar, konferensiyalar, seminarlar va Clarin tizimi bilan bog‘liq boshqa tadbirlar ro‘yxati mavjud. Shuningdek, tadbirlar o‘tkazilgan vaqti, joyi, tadbir nomi, shahri ham to‘liq ko‘rsatilgan. Foydalanuvchi istagan tadbir namoyishi va materiallari bilan tanishish va foydalanish imkoniyatiga ega.

Beshinchi bo‘limda (News) tizimdagi va tizim bilan bog‘liq yangiliklar va o‘zgarishlar, tizimga yangi qo‘shilgan lingvistik tahlillar natijasi, eng so‘nggi yangiliklar bayon etilgan.

Oltinchi bo‘limda (Contact) tizim bilan bog‘lanish manzillari joylashtirilgan. E-mail, telefon raqam, yangiliklar sayti, aloqa uchun saytlar ko‘rsatilgan. Ushbu tizimning “Resource families” qismida ko‘pgina til korpuslari jamlangan. Xususan, vazifasi va maqsadiga ko‘ra 15 ta korpus mavjud bo‘lib, ular o‘z ichiga yana bir qancha subkorpuslarni qamrab oladi (2-rasm).


Home / Language Resources / Resource Families

Resource Families

The CLARIN Resource Families provide a user-friendly overview per data type of the available language resources in the CLARIN infrastructure for researchers from the digital humanities, social sciences and human language technologies. The overviews are meant to facilitate comparative research and the listings are sorted by language.

The listings for each family include the most important metadata as well as brief descriptions, such as resource size, text sources, time periods, annotations and licences, as well as links to download pages and concordancers. In addition to the resources found in the CLARIN infrastructure, an overview of other existing valuable language resources, which have not yet been integrated into the infrastructure, is provided.

The listings also provide hyperlinks to other relevant materials, such as CLARIN workshops and tutorials, video lectures, and key publications.

If you would like to apply for funding for small projects that can help to extend the scope of the initiative, see Resource Families Project Funding.

Corpora
- Computer-Mediated Communication Corpora
- Corpora of Academic Texts
- Historical Corpora
- L2 Learner Corpora
- Legal Corpora
- Literary Corpora
- Manually Annotated Corpora
- Multimodal Corpora
- Newspaper Corpora
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources
- Language Models
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools
- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

The overviews were prepared by Darja Fišer and Jakob Lenardič and received funding from the European Union's Horizon 2020 research and innovation programme for

2-rasm. Clarindagi subkorpuslar

1. Computer-Mediated Communication Corpora. Kompyuter vositasidagi aloqa korpuslari. Kompyuter vositasidagi onlayn muloqotlarni o‘z ichiga oladi. Masalan, yangiliklar, saytdagi sharhlar, ijtimoiy tarmoqdagi ilovalar. Onlayn muloqotda ko‘pgina til o‘zgarishlariga uchraydi. Imloviy, ishoraviy xatoliklar kuzatiladi. Clarin tizimi 23 ta CMCni taqdim etadi. Ularning aksariyati, sloveniyaliklar uchun, shuningdek, golland, chex, fin, nemis, italyan, litva tillari uchun ham mavjud.

2. Corpora of Academic Texts. Akademik matnlar korpusi. Akademik matnlar ilmiy jurnallarda chop etilgan ilmiy maqola, tezis, konferensiya materiallari va monografiyalarni o‘z ichiga oladi. Ushbu tizimda 24 ta akademik matnlar korpusi mavjud, ulardan 2 tasi ko‘p tilli va 22 tasi bir tilli hisoblanadi.

3. Historical Corpora. Tarixiy korpus.

4. L2 Leaner Corpora. Ushbu turdagi korpusning o‘ziga xos xususiyatidan biri, yangi til o‘rganuvchilar o‘z xatolarini bilib borishlari mumkin.

5. Legal Corpora. Huquqiy korpus. Qonun hujjatlari, huquqiy hujjatlar, sud qarorlari va shunga oid materiallarni qamrab oladi.

6. Literary Corpora. Adabiy korpus. Adabiy korpus she’r va

nasriy asarlar, masalan, qissa, hikoya, roman, dramalarni o‘z ichiga oladi. Ushbu tizimda 45 ta adabiy korpusga kirish mumkin.

7. Manually Annotated Corpora. Izohli korpus. Lingvistik ma’lumotlarni qamrab olgan.

8. Multimodal Corpora. Multimediali korpus. Vidio, tasvir va yozuv ko‘rinishida ma’lumotlarni taqdim etadi.

9. Newspaper Corpora. Gazeta korpuslari. Gazeta to‘plamlari, ommaviy-axborot vositalari mavjud.

10. Oral History Corpora. Og‘zaki tarixiy korpus. Bir shaxs yoki ma’lum guruhning tarixi bilan bog‘liq bo‘lgan ma’lumotlarini to‘plash bilan shug‘ullanadi.

11. Parallel Corpora. Parallel korpus. Bunday korpuslar til o‘rganuvchilar uchun asosiy manba bo‘lib xizmat qiladi. Asosan, tarjima sohasida keng qo‘llaniladi.

12. Parliamentary Corpora. Parlament korpuslari.

13. Reference Corpora. Ma’lumotlar korpusi.

14. Sign Language Resources. Imo-ishoralar tili resurslari.

15. Spoken Corpora. Og‘zaki korpus. Bunda og‘zaki nutq, dialog, efirga uzatilgan ko‘rsatuvlarni o‘z ichiga oladi. Tilshunoslikda, xususan, dialektologiyada muhim manba bo‘lib xizmat qiladi.

Har bir korpusda korpus nomi, korpus tili va korpus haqida dastlabki umumiy ma’lumotlar berilgan bo‘lib, ularga kirish uchun havola va yuklab olish imkoniyati ham mavjud. Birgina og‘zaki korpus (faqat og‘zaki matnlardan tarkib topgan korpus [Hamroyeva, 2020. 34]) tarkibida 148 ta og‘zaki korpus mavjud bo‘lib, ulardan 134 tasida og‘zaki va matn ko‘rinishida, 14 tasida faqat transkripsiyasi mavjud. Korpuslarning aksariyati bir tilli bo‘lib, quyidagi 15 tilni qamrab oladi: Arab, chex, golland, eston, fin, fransuz, nemis, venger, italyan, nepal, norveg, polyak, sloven, ispan va shved. Aksariyat hollarda korpusni to‘g‘ridan-to‘g‘ri ma’lumotlar bazasidan yuklab olish yoki foydalanish uchun qulay onlayn qidiruv muhiti orqali foydalanish mumkin. Og‘zaki tilning korpuslari o‘z-o‘zidan yoki rejalashtirilgan nutqning transkripsiyalarini o‘z ichiga oladi, masalan, efirga uzatilgan yangiliklar yoki olingan rivoyatlar va dialoglar. Ular ko‘pincha ilova qilingan yozuvlar bilan izohlanadi. Ular fonologiya, suhbat tahlili va dialektologiya kabi turli xil lingvistik tadqiqotlar uchun qimmatli manbadir.

Og‘zaki korpusda bir tilning bir nechta korpuslari mavjud. Ular bir-biridan yaratilgan maqsadi va vazifasiga ko‘ra farqlanadi. Masalan ingliz tilining 5 ta korpusi kiritilgan bo‘lsa-da, ular mavzuviy jihatdan farqlanadi. 1. Aviadispetcherlar va uchuvchilar

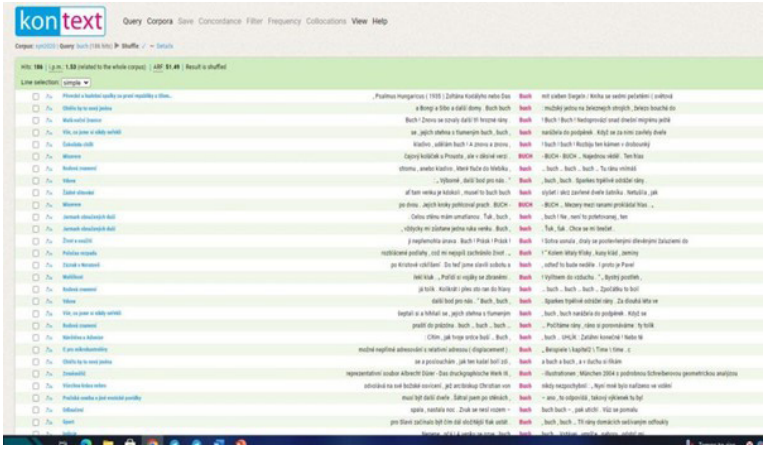
o‘rtasidagi aloqa yozuvlari mavjud. 2. Radio yangiliklaridan yozuvlar va matnlar mavjud. 3. Ushbu korpusda intervyu mavjud. 4. Yozilgan ma’ruzalar va seminarlar mavjud. 5. Ushbu korpusda ingliz tilidagi asosiy bog‘langan nutq jarayonlarini kiritish uchun mo‘ljallangan 460 ta qisqa jumlalar to‘plami mavjud. Ma’lumotlar bazasida audio fayllar, ovoz to‘lqin shakllari mavjud. Har bir til korpusida korpus turi haqida (Corpus), korpus tili (Language), korpus haqida umumiy ma’lumot (Description), yuklab olish bo‘limi (Availability) mavjud. Birgina chex tilining tizimda 3 xil korpusi joylangan.

1. Ko‘p bosqichli transkripsiyaga ega dialektal korpus. Hajmi 100000 so‘zdan iborat. Orfografik va fonetik (dialekt xususiyatlari) transkripsiyalangan, lemmatizatsiyalangan. Ushbu korpusda an’anaviy dialektologik material, asosan, monolog tipidagi nutqlar mavjud.

2. Norasmiy chex tilining korpusi (transkripsiya va audio). Hajmi 2,8 mln so‘zdan iborat. Ushbu korpus norasmiy suhbatlarni o‘z ichiga oladi.

3. Og‘zaki chex tili ma’lumotlar korpusi. Hajmi 770 000 token, 7 324 daqiqani tashkil etadi. Lemmatizatsiyalangan. Nutqlar, asosan dialog shaklda mavjud.

Dialekt korpusi butun Chexiya Respublikasida qo‘lga kiritilgan an’anaviy mintaqaviy dialektlarni taqdim etadi. Dialekt materiali Chexiya Respublikasining barcha dialektal mintaqalaridan kelgan ovoz yozuvlarini transkripsiya qilish orqali olingan. Korpus ikki darajadan iborat. Qadimgi dialektal daraja 1950-yillarning oxiridan 1980-yillargacha bo‘lgan davrda yozilgan yozuvlarni o‘z ichiga oladi. Yangi daraja 1990-yillardan hozirgi kungacha bo‘lgan davrni o‘z ichiga olgan ikkala qatlam uchun ham bizda bugungi kunda umuman uchramaydigan arxaik dialektal elementlarni o‘z ichiga olgan til ma’lumotlari mavjud. Dialekt korpusining ikkinchi versiyasi 220 000 dan ortiq so‘zlarni o‘z ichiga oladi. Birinchi dialektal korpusda 9745 ta, ikkinchi dialektal korpusda esa 43 628 ta nutq mavjud. Ushbu dialektal korpusda “buch” (kitob) so‘zi orqali qidiruv amalga oshirilganda quyidagi Kwic formatdagi natija olindi. Ushbu so‘z 186 ta matnda uchraydi (3-rasm).



3-rasm. Qidiruv natijasi ko‘rinishi

Ko‘rsatilgan barcha ma’lumotlar avtomatik tahlil orqali olinadi. Ma’lumotlarning aniqligi va ishonchliligi korpusdagi teglashning mukammalligiga bog‘liq. Shu sababli lemmatizatsiya, ya’ni asosiy so‘zshaklini ajratish va morfologik teglash muhim rol o‘ynaydi. Tizimdagi va korpusdagi ma’lumotlarda doimiy ravishda yangilanib, to‘ldirilib boriladi.

Xulosa

Olimlar korpusning maqsadi hamda oldiga qo‘yilgan vazifasiga qarab korpusning bir qancha tasnifini keltirishadi. Clarin tizimidagi korpuslar saqlash shakliga ko‘ra (ovozli, yozma, aralash) aralash; matn tiliga ko‘ra (bir tilli, ko‘p tilli) bir tilli, janriy mansubligiga ko‘ra (adabiy, dialektal, og‘zaki, publitsistik, aralash) aralash; korpusga kirish imkoniyatiga ko‘ra (erkin, tijorat korpuslari, yopiq) erkin; maqsadiga ko‘ra (tadqiqiy, illyustrativ) illyustrativ; dinamikasiga ko‘ra (dinamik, turg‘un) dinamik (chunki bu korpuslar yangilanib, to‘ldirilib boriladi); qo‘shimcha axborotga egaligiga ko‘ra (teglangan, teglanmagan) teglangan. Clarin tizimi mukammal korpuslar yig‘ilgan tizimdir. Unda nafaqat korpuslarni undan tashqari barcha raqamli til manbalari va vositalariga gumanitar va ijtimoiy fanlar tadqiqotchilarini qo‘llab-quvvatlash uchun yagona online muhit mavjudligi, keng qidiruv imkoniyatiga egaligi bilan ahamiyatlidir. Ma’lum bir tilda yaratilgan korpuslarning mukammalligi foydalanuvchiga ham, yangi til o‘rganuvchiga ham qulaylik yaratadi. Korpus qanchalik mukammal razmetkalangan bo‘lsa, uning qidiruv imkoniyati ham, olingan natija ham shunchalik aniq mukammal bo‘ladi.

Foydalanilgan adabiyotlar

Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. – Globe edit, 2020.

Hamroyeva Sh. Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati: Terminologik lug‘at. – Globe edit, 2020.

Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005.

Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2013.

https://www.clarin.eu/

https://www.clarin.eu/content/overview-clarin-centres

SPECIFICALLY ORAL CORPUSES IN THE CLARIN SYSTEM

Nilufar Muradova[1]

Abstract. The development of digital technologies is reflected in all areas. In particular, the issues of creating language corpora, natural language processing (NLP), machine translation are relevant in linguistics. Of course, these studies serve to increase the development and viability of our language. Mainly, language corpora are important in this. In world linguistics, large systems with excellent, extended search capabilities have been developed. One of these is the Clarin system. Clarin allows you to discover, explore, annotate, analyze, integrate and perform linguistic research on language data. Several language corpora are also included in this system. In this article, the purpose, mission, capabilities of the Clarin system; corpus, subcorpus, and verbal corpuses in the system are described. The types, possibilities and search system of oral corpora are described. In particular, general information about the spoken corpus of the Czech language, the working system of the corpus, the difference from the subcorpus, and search possibilities were studied.

Key words: *Clarin system, corpus, verbal corpus, lemmatization, search system.*

References

Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. – Globe edit, 2020.

Hamroyeva Sh. Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati: Terminologik lug‘at. – Globe edit, 2020.

Zaxarov V.P. Korpusnaya lingvistika. Uchebno-metodicheskoye posobiye. – Sankt-Peterburg, 2005.

Zaxarov V.P., Bogdanova S.Y. Korpusnaya lingvistika. – Irkutsk: IGLU, 2013.

https://www.clarin.eu/

https://www.clarin.eu/content/overview-clarin-centres

[1]*Muradova Nilufar Baxodir qizi* – Alisher Navoiy nomidagi o‘zbek tili va adabiyoti universiteti kompyuter lingvistikasi mutaxassisligi magistranti.
E-pochta: *muradovanilufar06@gmail.com*
ORCID: 0009000184741863

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 25.12.2023-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.