V SHO‘BA
KOMPYUTER LEKSIKOGRAFYASI: ONTOLOGIYA, TEZAURUS

KORPUSNING LINGVISTIK VA TA’LIMIY AHAMIYATI

Abduraxmonova Nilufar
O‘zbekiston Milliy universiteti
professori, filologiya fanlari doktori.
E-mail: n.abduraxmonova@nuu.uz

Askarova Umida
O‘zbekiston Milliy universiteti o‘qituvchisi.
E-mail: samina090822@gmail.com

Yulbarsov Ochilbek
O‘zbekiston Milliy universiteti o‘qituvchisi.

Annatatsiya: Ushbu maqolada xalqaro ta’limiy korpuslarning lingvistik instrumentariylari, lingvistik jihatdan aniqlangan xatolarni teglash tizimi, ta’limiy korpus loyihasi shakllantirilishning asosiy kategoriyalari, ta’limiy korpuslar toksonomiyasi kabi masalalar tahlilga tortilgan.
Kalit so‘zlar: TAGSELECT, ICEMARK, ICECUP, razmetkalash, kollokatsiyalar, subkorpus, transliteratsiya, token, korpuslar toksonomiyasi.

Annotation: In this article, the issue of linguistic instruments of international educational corpora, system of tags of linguistically identified errors, main categories in the formation of educational corpus project, taxonomy of educational corpora is highlighted.
Keywords: TAGSELECT, ICEMARK, ICECUP, sorting, collocations, subcorpus, transliteration, token, taxonomy of corpora.

Аннотация: В данной статье освещен вопрос лингвистического инструментария международных образовательных корпусов, системы меток лингвистически определяемых ошибок, основных категорий при формировании проекта учебного корпуса, а также вопрос таксономии учебных корпусов.
Ключевые слова: TAGSELECT, ICEMARK, ICECUP, сортировка, словосочетания, субкорпус, транслитерация, токен, таксономия корпусов.

Til va tilshunoslikni o‘rganilishida korpus manba sifatida foydalanish bo‘yicha omma qiziqishining ortib borishi ICAME (the international computer archive of modern and medieval English) a’zolari o‘rtasida bo‘lib o‘tgan muhokama natijasida 1994- yil TaLC yaratilishiga sabab bo‘ladi. Dastlabki ta’limiy korpusga oid konferensiya 1994- yilda Lankasterda bo‘lib, tadqiqotda kompyuter matn korpuslaridan foydalanish yaxshi yo‘lga qo‘yilgani va ular o‘qitish maqsadlarida keng foydalanib kelinyotganligini, korpus ma’lumotlaridan o‘quv materiallarini tayyorlashda

hamda xorijiy tillarni va lingvistik ta'limda talabalar tomonidan bevosita korpuslarning yaratilishiga doir muhokmalar kun tartibiga qo'yiladi.

Dastlabki ta'limiy korpuslarni yaratilishi Sylviane Granger va uning ilmiy jamoasi tomonidan "Ingliz tili o'rganuvchilari uchun xalqaro korpus" (ICLE – International corpus of learner English) yaratilgan. Ushbu korpusning maqsadi til o'rganuvchilarining umumiy yuzaga keluvchi xatolarini tasniflash orqali ularni tuzatish metodlarini yaratishga qaratilgan [D. Stewart. 2004:3].

Ushbu ta'limiy korpusni yaratishda ingliz tili o'rganuvchilarning quyidagi belgilari hisobga olingan [Sylviane Granger.1994:24]: ta'lim muhiti, yosh, ona tilisi, ta'lim bosqichi, vazifalarning tabiati.

Ushbu korpus uchun quyidagi kriteriyalari doirasida til o'rganuvchilar tanlab olingan:

1) o'rganuvchilar turi;
2) yoshi: yoshi kattalar (adults)
3) ta'lim bosqichi: yuqori daraja (advanced);
4) vazifasi: esselar

Ingliz tilining ta'limiy xalqaro korpusi ingliz tilining ikkinchi til sifatida franzus, nemis, xitoy, yapon, chex kabi millatga tegishli o'rganuchilar asos qilib olingan.

Ingliz tili xalqaro ta'limiy korpusi bir nechta lingvistik instrumentariylardan tarkib topgan.

- The ICE Markup Assistant – ushbu instrumentariy razmetkalash uchun standart belgilar to'plami asosida jarayonni soddalashtirilgan avtomat tizimi sanaladi;
- TAGSELECT – the Ice tag Selection system (teglash tizimi) korpusdagi har bir so'z uchun TOSCA teggeri yordamida generatsiya qilingan so'z guruhlari teglarini avtomatik saralash tizimi sanaladi;
- ICEMARK – TOSCA parseridan oldin teglangan matnni sintaktik razmetkalash uchun qo'llanadi. Shuningdek, boshqa funksiyalaridan biri so'z guruhlari teglarini o'zgartirish, matnda uchragan so'zlarni xatolari tahrirlash, teglanmagan gaplarni ko'rish kabi imkoniyatga ham egadir.
- ICECUP – korpus Utility dasturi ushbu instrumentariy korpusni tahlil qilish, konkordanslash va qidiruv kabu funksional jihatdan ahamiyatga ega.

XIX asrning oxiri va XX asrning boshlarida so'zning klaasik nazariyasi iki yirik leksik tadqiqotning yaratilishida katta rol o'ynadi. Roget tezaurus (1852) va Oksford ingliz lug'ati (Murray 1884-1928). Matnlarga qarab so'zlarning giponimi, sinonimi, antonimi va ularning ma'nolari leksik munosabatlarga ko'ra lug'atda beriladi. Har bir so'z bundan tashqari bir yoki bir nechta grammatik kategoriyalarga tegishli bo'ladi va ularning talaffuzi, etimologiyasi va tarixi beriladi. Ushbu nazariyaga ko'ra leksik birlik boshqa tizimlarga ham aloqador bo'lib, grammatika va fonologiya o'rtasida o'zaro munosabat yotadi. Sistem-funksional lingvistikaning markazida ham leksis turadi, zero bu modeldagi tizimlar leksik tanlov eng oxirgisi bo'lganidek ba'zan uni shu tarzda hosil qiladi [D. Stewart. 2004: 22]. XIX asrda yaratilgan Tezaurus lug'ati ta'sirida ko'pgina nazariy tadqiqotlar jonlana boshlaydi. XX asrda korpusga asoslangan o'quv lug'atlari Collins Cobuild, Oksford, Longman, Makmillan kabi lug'atlar yartailganligini guvohi bo'lamiz.

Ta'limiy korpuslar 1990-yillarda ommalashgan bo'lsa-da, biroq 1960-70-yillarda tillarni o'rganishda "xatoni tahlil qilish" (error analysis) tarzda metodologik yondashuv sifatida o'rganilgan.

Ma’lumki, so‘zlar muayyan darajada kombinatsiya hosil qilish uchun ma’lum turdagi matnlarda qo‘llanadi. Bu an’anaviy tilshunoslida so‘zlarning leksik ma’no qurshivini ko‘rsatishda matn, gap yoki kontekst jumlalari bilan qo‘llanganini ko‘ramiz. Endilikda so‘zni bir vaqtda turli maqsadlardan kelib chiqib tahlil qilish imkoniyatini yaratib beruvchi lingvistik instrumentariy sifatida korpusni ta’kidlash o‘rinli bo‘lar edi. So‘zning kollokatsiyalar bilan (so‘z birikma yoki turg‘un birikma) birgalikda uning tarkibida qo‘llanganda so‘zga doir u yoki bu ma’no anglashiladi. Demak ayrim so‘zlar uchun uning grammatik kategoriyaga xoslanishi gap tarkibida kelganda aniqlashadi: *o‘qish* (ot, harakat nomi yoki fe’lning buyruq shakli), *qil* (ot, fe’l), *yosh* (sifat, ot) kabilar.

Qachonki til foydalanuvchisi uchun biror so‘z semantik to‘plam yoki guruh bilan birgalikda kelganda semantik assotsiatsiyani aniqlash mumkin bo‘ladi [D. Stewart. 2004: 25].

Ta’limiy korpusga asoslangan til o‘qitish yoki o‘rganish jarayoni statistik yoki chastotaga asoslangaligi bilan ahamiyatlidir. Ayniqsa, u yoki bu so‘zning kontekstda birikma tarzda so‘z qurshovida kela olish imkoniyati so‘zlarning ko‘p ma’noli yoki omonim bo‘lgan holatda juda muhim ahamiyat kasb etadi.

Ta’limiy korpus til o‘rganuvchilar tomonidan tayyorlangan matnlarning tizimli kompyuterlashtirilgan majmuasi sifatida ta’rif keltiriladi [N.Nesselhauf. 2005:40]. Tizimli deganda, muayyan til o‘rganuvchilar guruhi sifatida tanlab olingan til bilish darajasiga ko‘ra jamlangan matnlardan tuzilgan korpus nazarda tutiladi. Ona tili korpusiga ixtisoslashgan korpusda matnning tabiiy yaratilganligi va tayyor olingan matnlar resurs bo‘lib hizmat qilishi bilan ajralib turadi. Ta’limiy korpusda xorijiy til yoki ikkinchi til sifatida o‘rganuvchilar uchun mo‘ljallangan bo‘lib, nazoratdagi til o‘rganuvchilar tomonidan yaratilgan chekli matnlardan tashkil topadi. Uar janr jihatdan esselar, og‘zaki intervyular bo‘lishi mumkin [D. Stewart. 2004: 40].

Rus tili ta’limiy korpusi rus tilida gapiruvchi turli millat vakillari tomonidan tayyorlangan materiallar asosida yaratilgan.

Korpusdagi matn va muallifga tegishli metama’lumotlar quyidagi ikki maydonga ajratilgan[7]:

I. Asosiy maydon:
- Og‘zaki / yozma matn
- Xorijiy / ona tili sifatida rus tili
- Muallifning dominant tili
- Rus tilini bilish darajasi

II. Ixtiyoriy maydon:
- Muallif jinsi
- Matn yaratilgan sanasi
- Janri

Ushbu korpus iqtisodiyot oliy maktab korpus texnologiyalari lingvistik laboratoriyasida xalqaro loyiha doirasida yaratilgan. Korpus 730000 tokendan iborat bo‘lib, matnlarning 56 foizi rus tili ikkinchi til sifatida so‘zlovchilar nutqidan, 44 foizi mahalliy (universitet va kollej talabalari) yuqori va o‘rta til bilish darajasiga ega rusiyzabon vakillar tomonidan olingan materiallarga asoslanilgan. Korpusning dastlabki versiyasi faqat Amerika ingliz tilida ona tilisi bo‘lgan loyihada cheklangan ishtirokchilar tomonidan yozib olingan. RULEC deb nomlangan subkorpus tarkibida akademik yozishmalardan iborat. Barcha ishtirokchilar ononim shaklda qatnashgan

[Abdurakhmonova, N.2019,2022]. Boshqa respondentlar ID kod bilan belgilangan [E.Rakhilina. 2016:66].

Lingvistik jihatdan aniqlangan xatolarni teglar tizimida ko‘rsatilgan:

Til sathi	TEG	Izohi
Imloviy xato	Graph	Lotin alifbosidan foydalanish
	Hyphen	Chiziqchani qo‘llash bilan xatolik
	Space	So‘zlar o‘rtasidagi bo‘shliqning tushib qolishi
	Translit	Atoqli otlarni transliteratsiya qilishdagi xatolik
	Ortho	Xato yozilgan harf
	Misspel	Bir tokendagi bir nechta o‘rinda yo‘l qo‘yilgan imloviy xato
Morfologik xato	Infl	Grammatik qo‘shimchlar xato qo‘llangan so‘z
	Deriv	Yasalgan so‘z
	Altern	So‘z o‘zagini qo‘llashdagi xato
	Num	So‘zdagi son kategoriyasini qo‘llash bilan bog‘liq xato
	Gender	Jins bilan bog‘liq xato
	Morhp	Boshqa morfologik xato
Sintaktik xato	AgrCase	Kelishik bilan bog‘liq xato
	AgrGender	Jinsning o‘zaro moslashish bilan bog‘liq xato
	AgrNum	Son bilan moslashishga doir xato
	AgrPers	Shaxs-son bilan bog‘liq xato
	Asp	Fe’l mayli bog‘liq xato
	Passive	Majhul nisbatni qo‘llash bilan bog‘liq xato
	Tense	Fe’l zamoni bilan bog‘liq xato
	Mode	Fe’lni qo‘llash bilan bog‘liq xato
	Refl	O‘zlik nisbatdagi fe’lni qo‘llash bilan bog‘liq xato
	Gov	Xato kelishik
	WO	So‘z tartibidagi xato
	Ref	O‘zlik olmoshi bilan bog‘liq xato
	Conj	Bog‘lovchini noto‘g‘ri qo‘llanishi
	Neg	Bo‘lishsizlikning xato qo‘llanishi
	Aux	Ko‘makchilarni xato qo‘llash
	Brev	Sifatlarning qisqa formasini qo‘llashdagi xato
	Syntax	Boshqa sintaktik xato
Tuzilish	Constr	Gap tuzilishfdagi xato
Leksik xato	Lex	Leksik xatolik
	CS	Kodlash bilan bog‘liq xato
	Par	Paronim bilan bog‘liq xato
	Idiom	Idiom bilan bog‘liq xato
Qo‘shimcha teglar	Del	So‘z yoki biror morfemada belgining tushib qolishi
	Insert	So‘z yoki biror morfemada belgining qo‘shilib qolishi
	Subst	So‘z yoki biror morfemada belgining almashib qolishi kelishi
	Transp	So‘z yoki biror morfemada belgining noo‘rin kelishi
	Transfer	Tildagi kelishikni transfer qilish
	Not-clear	Boshqa noma’lum fragmentlar

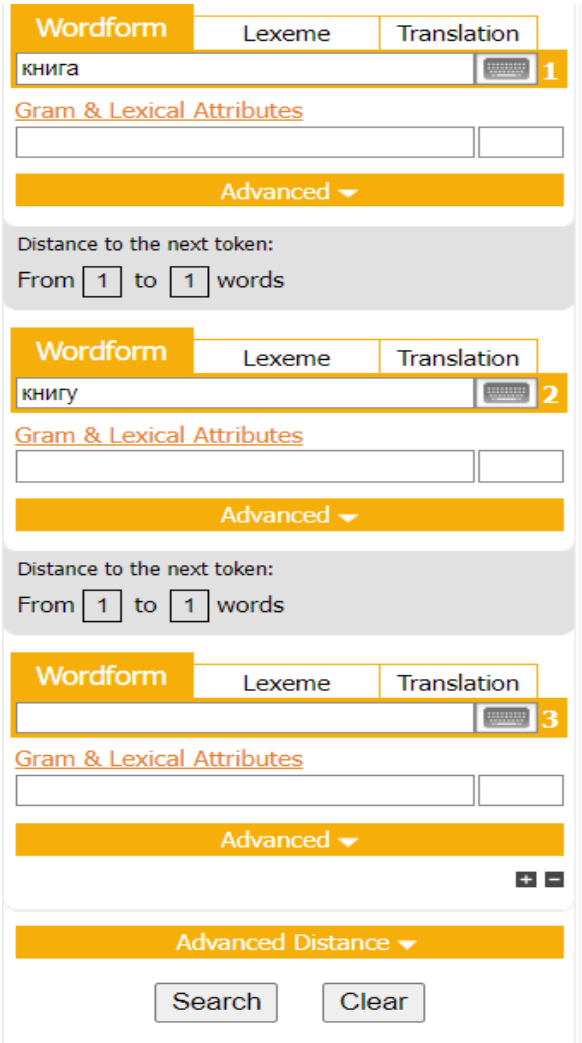
Rus tili ta’limiy korpusi 5 tilda gaplashuvchi (Amerika, franzus, koreys, qozoq va nemis tili) sodda, o‘rta va yuqori til bilish ko‘nikmasiga ega talabalardan olingan materiallardan tashkil topgan[Abdurakhmonova, N.2019,2022].

Y.Tononing fikricha, ta'limiy korpus loyihasi uch asosiy kategoriya asosida shakllantiriladi [Y.Tono.2003:800]: a) til bilan bog'liq kriteriya (janr, mavzu); b) mavzu bilan bog'liq kriteria (davomiy yoki yo'nalishlar kesimida; beixtiyoriy yoki tayyorlangan); c) o'rganuvchiga bog'liq kriteriya (ingliz tili xorijiy yoki ikkinchi til ekanligi (EFL/ESL), yosh, gender, ona tili, xorijiy tajriba).

Rus tili ta'limiy korpusi yuqoridagi me'yordan kelib chiqib, 8turdagi metama'lumotdan tarkib topgan 2 kategoriyaga biriktirilgan: muallifga tegishli hamda matnga tegishli. Muallifga tegishli belgilarga muallif IDsi, gender, til bilish darajasi, qaysi til vakili ekanligi. Muataxassislik darajasi CEFR (Common European Framework of Reference for Languages) va ACTFL (American Council on the Teaching of Foreign Languages) bo'yicha tanlanadi.

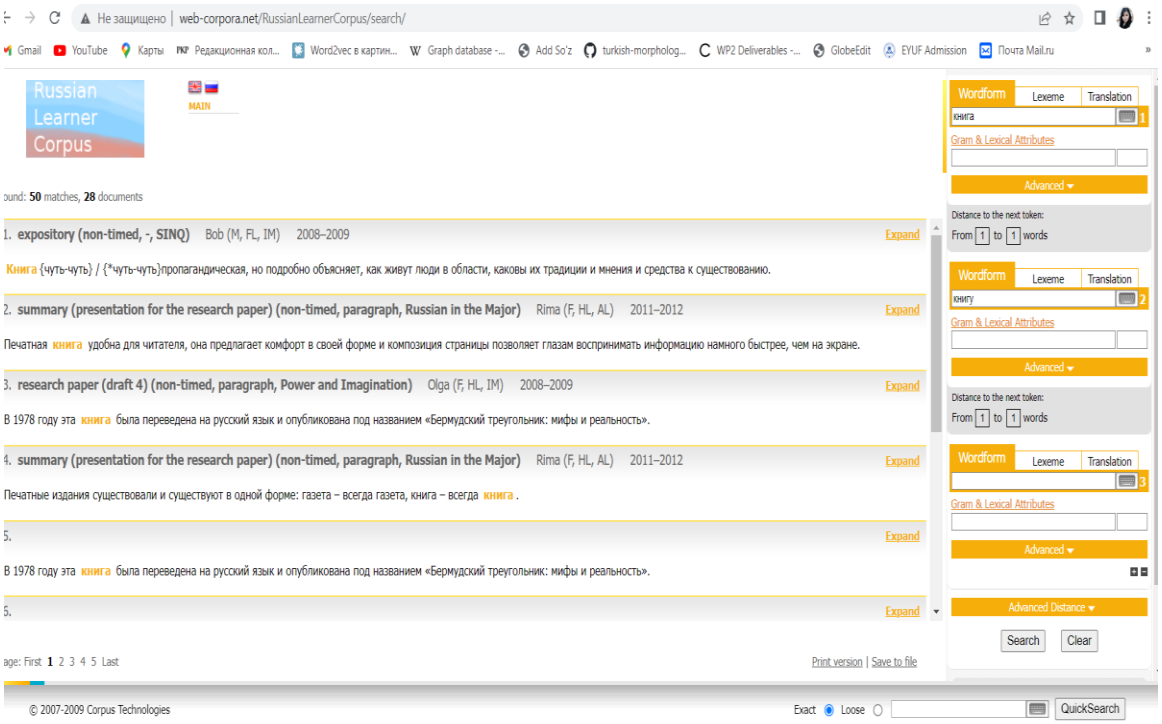
Y.Tononing fikriga ko'ra kamida ikki aspekt xatolarni belgilash annotatsiyalashdan iborat bo'lishi lozim. Birinchi guruh teglar lingvistik belgisiga ko'ra (qo'shimchalarni qo'shish, sintaktik moslashish) aniqlanaydi.

Rus tili ta'limiy korpusi annotatsiya instrumentariysi sifatida ochiq resursli JavaScript library Annotator.js ga asoslangan. Annotatsiya uch jihatga ega: 1-jihat asliyatdagi gapni ifodalash (annotator xatolarni belgilash), 2-jihat morfologik xatolar va imlolarni tuzatish orqali asliyatdagi matnni ko'rsatish, 3- ganing to'g'rilangan shaklini barcha sintaktik va leksik o'zgarishlar bilan borgalikda ko'rsatish.



2-rasm

Mazkur korpusning qidiruv tizimi quyidagi funksional imkoniyatlardan iborat: lemma, so‘z turkumlari va boshqa grammatik kategoriyalar (gender, son, nisbat, zamon).



3-rasm

Y.Tono ta’limiy korpuslar toksonomiyasi sifatida quyidagilarni keltiradi [Y.Tono. 2003:809]**:**

Proyekt	**Mavzusi/topshiriqlar** **Hajmi**	**Annotatsiya**
Yevropa:		
Xalqaro ingliz til ta’limiy korpusi **(ICLE** – international corpus of learner English)	- ingliz tili xorijiy til sifatida o‘qiydigan 3-4 kurs universitet talabalari -15 ta millat - Yozma esselar - 3 mln	-xato bo‘yicha teglash -POS (so‘z turkumi bo‘yicha teglash)
LINDSEI (og‘zaki ingliz tili Luven xalqaro ma’lumotlar bazasi)	-50 intervyu -ingliz tili xorijiy til sifatida o‘qiydigan 3-4 kurs universitet talabalari 100000so‘z	-orfografik
Longman ta’limiy korpusi **(LLC –** Longman Learners’ corpus)	-Barcha til darajalari -yozma esselar -10 mln	-POS- so‘z turkum bo‘yicha teglash -tijoriy maqsadlarga yo‘naltirilgan
Polyak-ingliz tili korpusi tadqiqi va ilovalari **PELCRA (**Polish-English Language Corpus Research and Applications**)**	-Barcha til darajalari -og‘zaki/yozma esselar Polyak -o‘rganuvchilar	POS bo‘yicha teglangan Mavjud emas
UAM korpusi	Talaba va o‘qituvchilarning tayyorlagan korpusi	

ISLE korpusi (ingliz tili xorijiy til bo‘lganlar uchun)	-20 minutlik nutq -o‘rta darajada ingliz tili darajasiga ega nemis va intalyan o‘rganuvchilar	-orfografik -urg‘u -ELRA orqali mavjud
JPU (Janus Pannonius University) korpusi	-universitet til o‘rganuvchilar Yozma materiallar	-oddiy matn
Kembridj ta’limiy korpusi **CLC (**Cambridge Learners Corpus**)**	-Barcha til daraja -10 mln.	-POS bo‘yicha teglangan -Xatolar bo‘yicha teglangan
IBLC (Indianapolis Business Learner Corpus**)**	-Biznes yozishmalar -oddiy matn	-oddiy matn -mavjud emas
OSIYO:		
JEFLL korpusi (Yaponiya)	-barcha darajalar uchun -og‘zaki va yozma nutq -1 mln.	-POS bo‘yicha teglangan -Xatolar bo‘yicha teglangan
Yaponiya o‘rganuvchilar uchun ingliz tili korpusi	-barcha darajalar uchun - yozma nutq -1 mln. Tarjimat	-oddiy matn -web orqali foydalanish mumkin
TAO /SST (Standart speaking test)	-barcha darajalar uchun - yozma nutq -1 mln. -15 min. intervyu	-Xatolar bo‘yicha teglangan (qisman)
TELEC talabalar korpusi	-Hong Kong o‘rganuvchilari -Universitet imtihonlaridagi yozuvlar -3 mln.	Oddiy matn
Poly U korpus	-universitet bitiruvchilari -Tezislar -282.000	Oddiy matn
NTOU korpus	-Ingliz tili xorijiy til sifatida o‘rganuvchilar -53.000	Oddiy matn
Ingliz tilini o‘rganuvchi yapon zabon vakillarining parallel korpusi	-ingliz va yapon tilidagi tarjimalar	Ma’lumotlar bazasi formatida
HKUST korpusi	Ingliz tili o‘rganuvchi xitoylik talabalar 10 mln. Yozma esselar va imtihon topshiriqlari	POS bo‘yicha teglangan

Foydalanilgan adabiyotlar:

1. D. Stewart, S. Bernardini, G. Aston Introduction: Ten years of TaLC / Corpora and language learners, Vol 17, John Benjamins Publishing Company Amsterdam/Philadelphia, 2004, - P. 3.
2. Sylviane Granger The learner corpus: a revolution in applied linguistics / English Today 39, Vol.10, No.3 (July 1994), Cambridge university press. – P. 26.
3. D. Stewart, S. Bernardini, G. Aston Introduction: Ten years of TaLC / Corpora and language learners, Vol 17, John Benjamins Publishing Company Amsterdam/Philadelphia, 2004, - P. 22.
4. Nadja Nesselhauf Collocations in a lerner corpus John Benjamins Publishing company, Amsterdam, 2005. – P. 40.

5. Abdurakhmonova, N., & Urdishev, K. (2019). Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL), 6(1-2019), 131-7.
6. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 73-75). IEEE.
7. Abdurakhmonova, N., Tuliyev, U., Ismailov, A., & Abduvahobo, G. (2022). UZBEK ELECTRONIC CORPUS AS A TOOL FOR LINGUISTIC ANALYSIS. In Компьютерная обработка тюркских языков. TURKLANG 2022 (pp. 231-240).
8. Abduraxmonova, N. Z. Q., & Urazaliyeva, M. Y. (2022). O ‘ZBEK TILI ELEKTRON KORPUSIDA (http://uzbekcorpus. uz/) OG ‘ZAKI MATNLAR KORPUSINI YARATISHNING NAZARIY VA AMALIY MASALALARI. Academic research in educational sciences, 3(3), 644-650.
9. Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. Building a learner corpus for Russian. In Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, pages 66–75, Umeå, Sweden. LiU Electronic Press.
10. Tono Y. Learner corpora: design, development and applications. In Proceedings of the 2003 Corpus Linguistics Conference, 2003. - P. 800–809.
11. http://web-corpora.net/RLC