UDK: 004;81'1

WEBCORPUS YONDASHUVI ASOSIDA INTERNETGA KORPUS SIFATIDAGI QARASHLAR

Ashirbaeva Madina Nuralievna
Chirchiq davlat pedagogika universiteti, Lingvistika va ingliz tilini o'qitish metodikasi kafedrasi o'qituvchisi va tayanch doktoranti
Mail: ashirbaevamadina5@gmail.com
Tel:+998998570691

Annotatsiya: *Ta'lim jarayonida zamonaviy texnologiyalar va resurslar muhim ahamiyatga ega. WebCorpus — internetda joylashgan katta hajmdagi matnlar to'plamidir. WebCorpus chet tillarini o'qitishda samarali vosita hisoblanadi. Bu resurslar til o'rganuvchilariga real hayotdagi tilni o'rganish (autentik), tahlil qilish va amaliyotda qo'llash imkoniyatini beradi. Ushbu maqolada keltirilgan yangi Internetga asoslangan yangi yondashuv o'qituvchilar va talabalar uchun qiziqarli va foydali bo'lishi mumkin. Undan tashqari maqolada WebCorpus yondashuvi, yoki Internetga korpus sifatidagi qarashlarning nazariy asoslari va korpus lingvistikasining rivojlanish bosqichlari va bu jarayonda "Web as Corpus"ning korpus lingvistikasiga ta'sirlari va uni birmuncha o'zgarishiga sabab bo'lganligini, korpusning qisqacha tarixi, unga berilgan ta'riflar bilan tanishtirib o'tadi.*

Kalit so'zlar*: "Web as Corpus"/WebCorpus yondashuvi, korpus lingvistikasi, Internet/Web, BNC, matnlar to'plami, WebCorpusning xususiyatlari, tillar reytingi, bog'langanlik.*

CORPUS-QUALITY VIEWS OF THE INTERNET BASED ON THE WEBCORPUS APPROACH

Abstract: *Modern technologies and resources are important in the educational process. Web as Corpus is a large collection of texts available on the Internet. It is an effective tool for teaching foreign languages. These resources allow language learners to analyze, learn and practice real-life (authentically) language. The new Internet-based approach presented in this article can be interesting and useful for teachers and students. In addition, the article describes the theoretical foundations of the Web as Corpus approach, or the Internet as a corpus, and the development stages of corpus linguistics, and the impact of "Web as Corpus" on corpus linguistics in this process and the reason for its slight change, a brief history of the corpus, introduces the definitions given to it.*

Keywords: *"Web as Corpus"/WebCorpus approach, corpus linguistics, Internet/Web, BNC, a collection of texts, features of WebCorpus, language rating, connectedness.*

КОРПУСНОЕ ПРЕДСТАВЛЕНИЕ ИНТЕРНЕТА НА ОСНОВЕ ПОДХОДА WEBCORPUS

Аннотация: *Современные технологии и ресурсы играют важную роль в образовательном процессе. Web as Corpus — это большая коллекция текстов, доступных в Интернете. Это эффективный инструмент обучения иностранным языкам. Эти ресурсы позволяют изучающим язык анализировать, изучать и практиковать реальный (аутентичный) язык. Новый интернет-подход, представленный в этой статье, может быть интересен и полезен учителям и студентам. Кроме того, в статье описаны теоретические основы подхода «web as Corpus», или «Интернет как корпус», и этапы развития корпусной лингвистики, а также влияние «Web as Corpus» на корпусную лингвистику в этом процессе и причины появления и изменение, краткая история корпусов, знакомит с данными ему определениями.*

Ключевые слова: *подход «Web as Corpus»/WebCorpus, корпусная лингвистика, Internet/Web, BNC, коллекция текстов, особенности WebCorpus, языковой рейтинг, связность.*

KIRISH

Korpus haqidagi to'liq ilmiy tadqiqot juda keng ko'lamli bo'lib, uni bir necha sahifalarda taqdim etish mumkin emas. Hozirda biz Internetning oltin asrida turibmiz, chunki u har qachongidan ham qulay va foydalanish uchun osondir. Internetning ushbu xususiyatlarini an'anaviy korpus bilan birlashtirsak, natijada "Web as Corpus" yondashuvi paydo bo'ladi. Nisbatan yangi yondashuv korpusning kelajagiga qanday ta'sir qilishini va Internetga asoslangan yangi korpus yondashuv an'anaviy korpusdan qanday farq qilishini ko'rib chiqadigan bo'lsak.

ADABIYOTLAR TAHLILI VA MUHOKAMALARI

Ushbu ishining tadqiqot obyekti sifatida korpus odatda "qo'lda tekshirishga imkon bermaydigan, ammo ma'lumotlarni miqdoriy va sifat jihatidan tahlil qilish uchun maxsus vositalardan foydalanishni talab qiladigan va o'z hajmiga ega baza" [Gatto,2013,7b] degan ta'riflarni Gatto taqdim etadi va bu texnologiyaning asosiy o'ziga xosligi bu uning har doim korpus lingvistikasi bilan bog'langanligidir. Kompyuterlarsiz so'z chastotalari, so'z shakllari va qiyosiy statistikasi bo'yicha muhim empirik tadqiqotlarni amalga oshirish imkoniyati deyarli mumkin emas.

1960-yillarning boshlarida, bir million so'zdan iborat Braun korpusi bilan kompyuterga asoslangan lingvistik tadqiqotlar boshlandi. 1980-yillarning boshlarida lingvist olim Sinkler bir qadam oldinga o'tib, sakkiz million so'zdan iborat korpus yaratdi. Vaqt o'tishi bilan texnologik muammolar va unga qo'yilagan cheklovlar

bartaraf bo'ldi va korpus lingvistikasining kamchiliklari ham birgalikda o'z yechimi topdi. Shu asnoda 1990-yillarning boshida Britaniya Milliy Korpusi (BNC) ishlab chiqildi va u yuz million so'zni o'zida topladi, bu avvalgi korpuslar bilan solishtirganda ilgari tasavvur ham qilib bo'lmaydigan ma'lumotlar bilan lingvistik tadqiqotlar olib borish imkonini berdi. Lekin bu yuz million so'zli korpusning ham o'z cheklovlari bor edi va bu empirik tadqiqotlar uchun hajm jihatidan yetarli emas edi, bu borada Hundt va boshqa olimlar "korpus lingvistikasining ba'zi sohalari uchun hatto BNC tipidagi yangi mega o'lchamli korpuslar hali ham mavjud emas" degan fikrlarni bildirdi.[Hundt, Nesselhauf & Blewer, 2007,1b].

O'sha vaqtga qadar eng katta korpus bo'lgan BNC joriy etilgandan 10 yil o'tgach, 1999 yilda boshqa bir olimlar Lourens va Giles o'sha paytda 800 million indekslanadigan veb-sahifalar mavjudligini taxmin qilishdi. Bugungi kunda ham 800 million sahifani tasavvur qilish qiyin, ularning har biri o'z mazmuni va matnlariga ega. Hozirgi kunda Internetda kamida 6 milliard sahifa borligi taxmin qilinmoqda. Ammo bu bilan miqdor sifatdan oshib ketmasligi lozim bo'lgan holatlar ham mavjud. Bu haqida Kilgarrif o'zining 'Web as Corpus" ya'ni "veb korpus sifatida" yondashuvini birinchi marta taqdim etgan maqolasida ta'kidlagan, "Internet bilan solishtirganda BNC ingliz qishlog'idagi bog'dir" degan edi [Kilgarriff,2001,342]. Ko'p sonli materiallarni o'z ichiga olgan Internetning keng xilma-xilligi avval aytib o'tilganidek ikki qirrali qilich bo'lishi mumkin.

Kilgarrif "Web as Corpus" yondashuviga uchta qarshi argumentlarni ham taqdim etdi, bu kamchiliklarni yondashuv avtori ham tan olgan. Aniqlangan birinchi muammo shundaki, hamma hujjatlar ham matn ko'rinishida bo'la olmaydi, sahifalar videolar va tasvirlardan tuzilgan bo'lishi mumkin. Ikkinchi qarshi argument Internet doimiy ravishda o'zgarib turadi, shuning uchun aniq chegaralarni belgilash qiyin bo'lishi mumkin. Uchinchi qarshi dalil Internetda dublikatlar bo'lishi va bir nechta tilni o'z ichiga olishi mumkin bo'lgan sahifalar bo'lishi mumkin, bu kamchilik tadqiqotning yakuniy natijalarini o'zgartirishi mumkin. [Kilgarriff,2001].

Korpus lingvistikasi BNCga juda ko'p qarzdor, u o'z zamonida tadqiqotchilar uchun koplab yo'nalishlarni ochib bergan korpus edi. Ammo hozir dunyoda o'sha davrda bo'lmagan Internet yoki "Web" deb nomlanuvchi vosita mavjud. Ushbu Web

deb nomlanuvchi virtual konstruksiya o'z foydalanuvchilariga turli sohalardagi ko'p sonli turli veb-sahifalarga masofadan turib sichqonchaning bir marta bosishi orqali kirishni va bepul foydalanish imkonini beradi. Kilgarrif ta'kidlaganidek: "BNC va shu kabi sobit korpuslar katta qiymatga egaligi yoqolmagan bo'lsa ham, tilning tabiati haqidagi eng provokatsion savollarni taqdim etuvchi Internetdir".[Kilgarriff, 2001, 344]. Internetning eng muhim xususiyatlaridan biri bu uning "bog'langanligi" [Schafer & Bildhauer, 2013,8 b], va bu fikrni tasdiqlovchi dalillarni topish qiyin emas, uning nomi "world wide web" deb nomlanadi.

Tarixan, anglafon (anglaphone) tarkibi internetda topilgan tarkib jihatidan boshqa tillarga soya bo'lishi ajablanarli emas. Kontent miqdori foydalanuvchilar soniga bog'liq emasligini inobatga olish muhim, buni 2-jadvalda, Internetda eng ko'p foydalaniladigan tillar reytingini 1-jadvalda ko'rish mumkin.

Top ten languages used in the web-2017
(number of Internet users by language)

Top ten languages in the Internet	World population for this language	Internet users by language	Internet penetration (%Population)	Internet users growth	Internet users % of World participation
English	1,462,008,909	1,055,272,930	72,2 %	649,7%	25,4%
Chinese	1,452,593,223	804,634,814	55,4%	2,390,9 %	19,3%
Spanish	515,759,912	337,892,295	65,5%	1,758,5 %	8,1%
Arabic	435,636,462	219,041,264	50,3%	8,616,0 %	5,3%
Portuguese	286,455,543	169,157,589	59,1%	2,132,8 %	4,1%

Indonesian	299,271,514	168,755,091	56,4%	2,845,1%	4,1%
French	412,394,497	134,088,952	32,5%	1,017,6%	3,2%
Japanese	127,185,332	118,629,672	93,3%	152,0%	2,9%
Russian	143,964,709	109,552,842	76,1%	3,434,0%	2,6%
German	96,820,909	92,099,951	95,1%	234,7%	2,2%
Top 10 languages	5,135,270,101	3,209,122,400	62,5%	1,091,9%	77,2%
Rest of the languages	2,499,488,327	950,318,284	38,0%	935,8%	22,8%
World total	7,634,758,428	4,159,440,684	54,5%	1,052,2%	100,0%

Jadval 1. Internetda eng ko'p foydalaniladigan tillar reytingi.

Ingliz tili har doim Internetning "lingua franca"si bo'lib kelgan, ammo bu uning asosiy til ekanligini anglatmaydi. Sababi Internetda berilgan sahifalarning uchdan ikki qismidan ko'prog'i ingliz tilida berilgan [Grefenstette & Nioche, 2000].

Birinchi jadvaldan ko'rinib turibdiki, ingliz tilini bilmaydigan foydalanuvchilarning o'sishi ingliz tiliga qaraganda ancha yuqori. Bu globallashuv va Internetning ko'p tilliligi ta'sirining natijasidir. Jadvalga ko'ra, eng ko'p ishlatiladigan tillar ingliz va xitoy tillari bo'lib, boshqa barcha tillardan aniq ustunligini ko'rish mumkin. Xitoylik foydalanuvchilarning haqiqiy sonini va ularning o'sish sur'atlarini hisobga olsak, kelajakda ingliz tilidan oshib ketishi ajablanarli emas. Aytib o'tish kerak bo'lgan yana bir ma'lumot arab tilida foydalanuvchilarrning o'sib borishi, bu Internetda rivojlanayotgan til bo'lib, o'sish sur'atlari boshqa tillarga qaraganda yuqori. Internetning til muhitidagi o'zgarishlar, umuman olganda dunyo jamiyatidagi o'zgarishlarning ifodasidir.

Veb ko'p tilli muhit bo'lib hizmat qiladi, u barcha turdagi tillardan foydalanishni taklif eta oladi. "Internet eklektik vositadir va bu uning ko'p tilli

inklyuzivligida ham ko'rinadi. U nafaqat til ichidagi barcha lingvistik uslublar uchun makon taklif etadi, balki barcha tillar uchun makon bo'la oladi, agar undan foydalanish uchun imkon bo'lsa bas" [Crystal 2006,229 b]. Bundan ko'rinadiki, tillarning tarqalishida iqtisodiy omillar muhim rol oynaydi, chunki kambag'al mamlakatlarda Internetdan foydalanish imkoniyati kamroq. Iqtisodiy tafovutdan tashqari, Internetdan teng foydalanishga to'sqinlik qiladigan yana bir noqulaylik bu lotin alifbosi uchun maxsus ishlab chiqilgan tizim va bu lotin tiliga asoslanmagan alifbolarni kodlash uni yanada qiyinlashtiradi [Crystal, 2006; Gatto, 2013,53 b]. Quyidagi 2 jadvalda internet dunyo aholi orasidagi geografik taqsimoti ko'rsatilgan.

World Internet usage and population statistics (2018)						
World regions	Population	Population % of World	Internet users	Penetration rate (%Pop.)	Growth 2000-2018	Internet users %
Africa	1,287,914,329	16,9%	464,923,169	36,1%	10,199%	11,0%
Asia	4,207,588,157	55,1%	2,062,197,366	49,0%	1,704%	49,0%
Europe	827,650,849	10,8%	705,064,923	85,2%	57,0%	16,8%
Latin America	652,047,996	8,5%	438,248,446	67,2%	2,325%	10,4%
Middle East	254,438,981	3,3%	164,037,259	64,5%	4,894%	3,9%
North America	363,844,662	4,8%	345,660,847	95,0%	21,9%	8,2%
Australia	41,273,454	0,8%	28,439,277	68,9%	27,3%	0,7%
World Total	7,634,758,428	100,0%	4,208,571,287	55,1%	1,066%	100,0%

Jadval 2. Jahon Internetdan foydalanish va aholi statistikasi.

Biz "Web" deb aniqlagan ma'lumotlarning ushbu birikmasining bir qismi bo'lgan har bir element boshqa element bilan bog'langan, barcha ma'lumotlar magistrallari orasida izolyatsiya qilingan elementni topish juda kam uchraydi. Internetning asosiy xususiyatlaridan yana biri uning doimiy kengayish xususiyatidir. Veb har soniyada o'sib boradi, doimiy ravishda har qanday shakl va ko'rinishdagi yangi ma'lumotlarga ega bo'ladi, Internetni har daqiqada eksponent ravishda kengaytiradi deyish unchalik to'g'ri bo'lmaydi. Gap shundaki, hozirgi vaqtda Internetning hajmi va o'sishiga shubha qilish mumkin emas. Internetning o'zaro bog'liqligi va ulkanligi haqidagi dalillardan foydalanish allaqachon 'Web as Corpus" yondashuvini himoya qilish uchun juda to'g'ri bo'lar edi. Agar bu bog'liqlik va uning doimiy o'sishi birlashtirilsa, Internet foydalanuvchilar uchun bebaho lingvistik ta'minotchiga aylanadi, chunki endi ular deyarli cheksiz imkoniyatlarni taqdim etuvchi vositadan foydalanishlari mumkin.

Kilgarrif va Grefenstette (2003,1 b) "Introduction to the special issue to the Internet as Corpus" nomli kirish maqolasida ta'kidlaganidek, Internetdan to'g'ri foydalanilsa "... tilshunoslarning ajoyib o'yin maydonchasi"dir degan edi.

Shu o'rinda Kilgarrif va Grafenstette (2003) tomonidan berilgan "korpus til yoki adabiy tadqiqot obyekti sifatida qaralganda korpus matnlar to'plami" degan ta'rifga qaytsak. Agar biz ushbu oddiy ta'rifni hisobga olib va Schafer va Bildhauer (2013) tomonidan taklif qilingan Internetning "bog'langanlik" xususiyati bilan birlashtirsak, bu ikki g'oyaning kombinatsiyasi o'zaro qiziqarli ta'sir qilish mumkin. Internetning bir-biriga bog'langan obyekt ekanligini tan olsak, bu uni yagona birlik yoki turli sohalar haqidagi yagona "matnlar to'plami" sifatida ko'rib chiqish mumkinligini anglatadi, ammo shunga qaramay to'plam Kilgarrif va Grafenstette ta'rifining bir qismi.

Endi Internetni ushbu matnlar to'plami sifatida tan olinganidan so'ng, Kilgarrif va Grafenstette ta'rifining ikkinchi qismiga quyidagi assotsatsiyani oddiy formula yordamida ko'rib chiqaylik:

a. "Korpus til yoki adabiy tadqiqot obyekti sifatida qaralganda matnlar to'plamidir"

b. Web—o'rganish obyekti bo'lishi mumkin bo'lgan matnlar to'plami.

Agar ikkala taklif ham to'g'ri deb hisoblansa, unda ulardan xulosa chiqarish mumkin: "Internet, to'g'ri foydalanilsa, haqiqatan ham korpus". Ko'pgina olimlarning fikriga tayanib va tadqiqot uchun Internetni "(...) tilni o'rganish obyekti sifatida" ko'rib chiqish lozim, Internet haqiqatan ham korpus ekanligiga da'vo qilish mumkin.

Amaliy tomondan, foydalanuvchi biron bir qidiruv tizimida so'z yoki iborani qidirganda, foydalanuvchi vebdan huddi korpus kabi foydalansa bo'ladi. Lekin "Web as Corpus" bilan bir hil jarayon emas, chunki korpus tilshunosligida Internetdan foydalanishning turli usullari mavjud. "Web as Corpus" yondashuvi turli mualliflarning nazariya va takliflarining uyg'unligi bo'lsa, unda uni korpus lingvistikasi tadqiqotlarida amalda sinab ko'rishning turli usullari ham mavjud. Shunday usullarning biri—bu qidiruv tizimlari kabi har qanday foydalanuvchiga taqdim etadigan to'g'ridan-to'g'ri vositalar yordamida Internetdan to'g'ridan-to'g'ri foydalanish. Bu usulni birinchilardan bo'lib Kilgarrif taqdim etgan va eng keng tarqalgan usullardan biridir.

Ammo tadqiqotchilar orasida yana bir tendensiya mavjud bo'lib, u "Web as Corpus" yondashuviga tegishli bo'lib, undan foydalanib kelinmoqda va u Internetdan ulkan korpus sifatida to'g'ridan to'g'ri foydalanish emas, balki "yuklab olinadigan va qayta ishlanadigan matnli ma'lumotlar manbai sifatida tadqiqotlarda foydalanishdan iborat" [Ferraresi, 2009,2b]. Yoki boshqacha qilib aytganda, internet korpusning o'zi emas, balki katta offlayn korpuslarni yig'ish manbai sifatida qarash [Hundt, 2009,2]. Bu ikki xil nuqtai nazarni ajratib turuvchi chiziq esa nozikdir.

Internetni korpus sifatida ko'rib chiqishning bunday talqinining qoralovchilari ham mavjud, chunki Sinkler "veb korpus emas, chunki uning o'lchamlari noma'lum va doimiy ravishda o'zgarib turadi va u lingvistik nuqtai nazardan ishlab chiqilmagan" [Sinclair,2005].

XULOSA

Shunday qilib ushbu maqolada, Internetni korpus sifatidagi qarashlarning, ya'ni "Web as Corpus" yondashuvining asosiy tushunchalari kiritib o'tildi. Bunda umumiy korpus lingvistikasining rivojlanish bosqichlari, Internet va kompyuterizatsiya jarayonining korpus lingvistikasiga va uning rivojiga ta'sirlari ko'rib chiqildi. Bu bilan "Web as Corpus"ni korpus lingvistikasining kelajagi degan

xulosalarga kelishimiz mumkin bo'ladi. Va kundan-kunga qo'shilayotgan va o'zgarayotgan barcha yangi qadamlar va xususiyatlar korpus tilshunosligi rivoji uchun qiziqarli yangi istiqbollarni ochadi. "Web as Corpus" yondashuvi evolyutsiya zanjiridagi yangi bir qadamdir.

FOYDALANILGAN ADABIYOTLAR:

1. Ferraresi, A (2013). Google and Beyond: Web-As-Corpus Methodologies for Translators, Revista Tradumàtica,7, Available at: http://www.fti.uab.cat/tradumatica/revista/num7/articles/04/04art.htm
2. Gatto, M. (2013). 'The Web as Corpus: Theory and Practice' Cambridge: Cambridge University Press.
3. Grefenstette, G. and Nioche J. (2000), 'Estimation of English and non-English language use on the WWW'. In Proc. RIAO (Recherche d'Informations Assisté par Ordinateur), pp.237–246.
4. Hundt M., Nesselhauf, N. and Biewer, C. (eds) (2007), 'Corpus Linguistics and the Web'. Amsterdam.
5. Kilgarriff, A. (2001), 'Web as Corpus', in Proceedings of the Corpus Linguistics Conference (CL 2001), University Centre for Computer Research on Language Technical Paper, Vol. 13, Special Issue, Lancaster University.
6. Kilgarriff, A. and Grefenstette, G. (2003), 'Introduction to the Special Issue on the Web as Corpus', in Computational Linguistics, 29, 3, pp.33–47.
7. Schafer, R. and Bildhauer, F. (2013). 'WebCorpus Construction'. San Francisco: Morgan & Claypool.
8. Sinclair, J. 2005), 'Corpus and Text. Basic Principles', in Wynne, M. (ed.), Developing Linguistic Corpora: A Guide to Good Practice, Oxford: Oxbow Books, 1–16. http://ahds.ac.uk/linguistic-corpora.