BROWN KORPUSI VA 1-AVLOD INGLIZ KORPUSLARI HAQIDA

Ataboyev Nozimjon Bobojonovich
F.f.f.d.(PhD), dotsent
Yusupova Ra`no Davronovna
Buxoro Davlat Universiteti, Xorijiy tillar fakulteti,
Ingliz adabiyotshunosligi yo`nalishi
2-bosqich magistranti

Annotatsiya: Ingliz tilini mukammal o`rganish va lingvistik jihatdan tahlil qilish uchun ingliz tili korpuslarini chuqur o`rganib tahlil qilish zarur. Brown korpusi va shunga o`xshash til korpuslari tilning barcha xususiyatlaridan tortib, uning o`rganilishida qilinayotgan ilmiy izlanishlar va yangiliklarning ba`zilari bilan tanishib chiqilgan. Birinchi va ikkinchi avlod korpuslari va ularning yaratilishi haqida ma`lumotlar keltirilgan.

**Kalit so`zlar:** til korpusi, Brown korpusi, teglash, dasturlash, so`z chastotasi, korpus versiyalari.

Kirish

Ingliz tili tilshunosligida hozirgi kunga qadar tilni mukammal o`rganish maqsadida yaratilib kelinayotgan til korpuslariga duch kelamiz. Lekin ushbu korpus lingvistikasining fan sohasidagi tarixi nisbatan qisqadir. Birinchi zamonaviy kompyuterlashgan korpus 1961-1964 yillarda Brown universitetida ikki tilshunos olim Genri Kuchera va Nelson Frensis tomonidan yaratilgan degan umumiy tushuncha mavjud. Ya`ni ular o`zlarining klassik asarlari "*Computational Analysis of Present-Day American English"(Hozirgi amerika ingliz tilining kompyuterlashtirilgan tahlili)*ni nashr etishdi va u jahonga statistik ma`lumotlarni taqdim etdi.[1] Turli xil kompyuterlashtirilgan yoki dasturlangan tekshiruvlar uchun Kuchera va Frensis tilshunoslik, psixologiya, statistika va sotsiologiya elementlarini o`zida mujassam etgan keng qamrovli va turli-tuman bo`lgan opusni tuzdilar. Birinchi leksikostatistik tahlil nashr etilgandan ko`p o`tmay, Boston nashriyoti Houghton-Mifflin o`zining yangi *Amerika Merosi Lug`ati* uchun bir million so`z, uch qatorli iqtiboslar bazasini taqdim etish uchun Kucheraga murojaat qildi. 1969 yilda paydo bo`lgan ushbu yangi lug`at so`z chastotasi uchun korpus lingvistikasidan foydalanilgan holda tuzilgan birinchi lug`at edi.[2]

Dastlabki Brown korpusida faqat so'zlarning o'zi va har biri uchun joylashuvni aniqlovchi to`plam mavjud edi. Keyingi bir necha yil ichida gap bo`laklari teglari qo'llanildi. Greene va Rubinning yorliqlash dasturi bunga katta yordam berdi, ammo keng qamrovli qo`llanma talab qilindi.

Mavzuga oid adabiyotlarning tahlili. Brown korpus taxminan 80 ga yaqin gap

[1] Francis, W. Nelson & Henry Kucera. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press. 1967.

[2] Winthrop Nelson Francis and Henry Kučera. Frequency Analysis of English Usage: Lexicon and Grammar, Houghton Mifflin. 1983.

bo`laklari to`plamidan iborat bo`lib, qolaversa, qo`shma gaplar, qisqartmalar, xorijiy so'zlar va boshqa bir nechta holatlar uchun maxsus ko'rsatkichlardan foydalanilgan va Lancaster-Oslo-Bergen korpusi kabi ko'plab keyingi korpuslar uchun namuna bo'la oldi. Korpusni teglash, Endryu Makki tomonidan dasturlashtirilgan va ingliz tili grammatikasiga oid kitoblarda yozilgani kabi ancha murakkab statistik tahlilni amalga oshirish imkonini berdi.

Qiziqarli natijalardan biri shundaki, hatto juda katta namunalar uchun ham so'zlarni paydo bo'lish chastotasini kamaytirish tartibida grafik qilish giperbolani ko'rsatadi: eng ko'p uchraydigan n-chi so'zining chastotasi taxminan 1/n ga proportsionaldir. Shunday qilib, "the" so`zi Brown korpusining qariyb 7% ni tashkil qiladi, "to" va “of” esa 3% dan ko'proq; 50 000 ga yaqin so'zlarning taxminan yarmi hapax legomena: korpusda faqat bir marta uchraydigan so'zlardir. Bu oddiy daraja-chastota bog`likligi Jorj Kingsli Zipf tomonidan favqulodda xilma-xil hodisalar ro`y berishi uchun qayd etilgan va Zipf qonuni sifatida tanilgan.

Brown korpus_ korpus tilshunosligi sohasida ilk bor yaratilgan bo'lsa-da, hozirgi vaqtda odatiy korpuslar (masalan, zamonaviy Amerika inglizlari korpusi, Britaniya milliy korpusi yoki ingliz tilining xalqaro korpusi) taxminan 100 million so'zdan iborat bo'lib, hajm jihatdan ancha kattaroqdir.

Brown korpusi zamonaviy korpusshunoslikni shakllantirgan desak mubolag`a bo`lmaydi. U nafaqat yangi ingliz tili, balki barcha zamonaviy milliy korporatsiyalar uchun namuna bo`lib xizmat qiladi va hanuzgacha turli tadqiqotlarda ma`lumotlar to`plami sifatida foydalanilmoqda. Yuqorida aytib o`tganimdek, 1961-1964 yillarda qo`shma shtatlarda nashr etilgan asarlardan jamlangan bo`lib, taxminan bir million so`z bo`lib, ingliz tilining 500 ta namunasini o`z ichiga olgan umumiy til korpusidir.[3]

Tadqiqot metodologiyasi. Qayta ko'rib chiqilgan nashr 1971 yilda chiqarilgan bo'lib, asosan yetti yil foydalanishda paydo bo'lgan matn haqidagi ma'lumotlarni o'z ichiga oladi. Hozirgi tahrir yanada kengroqdir, chunki u korpusning yaqinda tayyorlangan versiyalari, xususan, 1979 yilda Braun universitetida to'ldirilgan "yorliqli" matn haqida ma'lumotni o'z ichiga oladi. Korpusning ikkita to'liq o'qilishi natijasida unga ikki xil tuzatishlar kiritildi: tayyorlashdagi xatolar yaqinda nashr etilgan nusxalarda tuzatilgan asl lenta va 33-176-betlardagi alohida namunalar tavsifida qayd etilgan asosiy matndagi keyingi tipografik xatolar va anomaliyalar(boshqa holatlar) .

Korpusning muqobil versiyalarini ishlab chiqarishda xizmat qilgan tilshunos-olimlar_Gerald M. Rubin, Barbara Grin Levin, Sandra Pirs, Patrisiya Strauss, Stiven Rits, Endryu Makki, Jostein Xauge va Donald Sherman va boshqalardir. Yozish vaqtida Korpusning 160 dan ortiq nusxalari muomalada bo'lib, korpusdan foydalangan holda yoki unga havola qilingan nashr etilgan asarlarning yaqinda bibliografiyasi 57 ta bandni o'z ichiga oladi.

Hozirgi amerika ingliz tilining ushbu standart korpusi 1961 yil taqvim yilida Qo'shma Shtatlarda chop etilgan tahrirlangan ingliz nashrining 1 014 312 so'zlik matnidan iborat. Garchi barcha materiallar birinchi marta 1961 yilda nashr etilgan bo'lsa-da, ularning ba'zilari, shubhasiz, avvalroq yozilgan. Biroq, avvalgi matnning

[3] Kirsten Malmkjær, *The Linguistics Encyclopedia*, 2nd ed, Routledge, 2002, ISBN 0-415-22210-9, p. 87.

ikkinchi nashri yoki qayta nashri ekanligi ma'lum bo'lgan hech qanday material kiritilmagan.

Korpus har biri 2000+ so‘zdan iborat 500 ta namunaga bo‘lingan. Har bir namuna gap boshida boshlanadi, lekin xatboshi yoki boshqa kattaroq bo‘linish shart emas va har biri 2000 so‘zdan keyin birinchi jumla oxirida tugaydi. Namunalar nasrning keng ko‘lamli uslublari va turlarini ifodalaydi. Nasrdan farqli ravishda o'ziga xos lingvistik muammolarni ko'rsatganligi sababli bayt kiritilmagan. (Ammo nasr namunalarida keltirilgan qisqa misralar saqlanadi.) Masalan, drama haqiqiy yozma nutq emas, balki og'zaki nutqning xayoliy rekreatsiyasi sifatida chiqarib tashlandi. Badiiy adabiyot kiritilgan, ammo 50% dan ortiq dialogdan iborat namunalar qabul qilinmagan. Namunalar sub'ektiv ravishda aniqlangan mukammallik uchun emas, balki ularning vakillik sifati uchun tanlangan. Korpus sarlavhasida standart so'zining ishlatilishi hech qanday tarzda uning "standart inglizcha" sifatida ilgari surilganligini anglatmaydi; u shunchaki ushbu korpusdan bir xil ma'lumotlardan foydalanish muhim bo'lgan qiyosiy tadqiqotlar uchun ishlatilishiga umid bildiradi. Ma'lumotlarni tayyorlash va kiritish kompyuter ishidagi asosiy muammo bo'lganligi sababli, maqsad standartlashtirilgan formatda jiddiy o'lchamdagi ehtiyotkorlik bilan tanlangan va tayyorlangan materialni taqdim etish edi. Korpus qo'shimcha ma'lumotlarning ingliz yoki boshqa tillarda tayyorlanishi va taqdim etilishi uchun namunani belgilashda standart bo'lishi mumkin.

Bu jarayon ikki bosqichdan iborat: dastlabki sub'ektiv tasniflash va har bir kategoriyadagi qancha namunalardan foydalanish to'g'risida qaror qabul qilish, so'ngra har bir toifadagi haqiqiy namunalarni tasodifiy tanlash. Ko'pgina toifalarda Braun universiteti kutubxonasi va Providens Athenaeum xoldingi tasodifiy tanlovlar qilingan koinot sifatida ko'rib chiqildi.[4] Ammo ba'zi toifalar uchun bu ikkita katakchadan tashqariga chiqish kerak edi. Kundalik matbuot uchun, masalan, Nyu-York jamoat kutubxonasida mikrofilm fayllari saqlanadigan Amerika gazetalari ro'yxati ishlatilgan (Providence Journal qo'shilishi bilan). Asosan materiallarning ayrim toifalari o'zboshimchalik bilan qaror qabul qilishni talab qildi; "Mahoratlar va sevimli mashg'ulotlar" va "Ommaviy ilm" toifalaridagi ba'zi davriy nashrlar Nyu-York shahridagi eng yirik ikkinchi qo'l jurnal do'konlaridan birining mazmunidan tanlangan.

Tahlil va natijalar. Asosiy toifalar va ularning bo'linmalari ro'yxati 1963 yil fevral oyida Braun universitetida bo'lib o'tgan konferentsiyada tuzilgan. Konferentsiya ishtirokchilari har bir toifadagi namunalar soni bo'yicha ham mustaqil ravishda o'z fikrlarini bildirishdi. Foydalanilgan raqamlarning dastlabki to'plamini olish uchun bu raqamlar o'rtacha hisoblangan. Keyinchalik tanlovlarni amalga oshirishda to'plangan tajriba asosida bir nechta o'zgarishlar kiritildi. Yaxshiroq bo'linish 1961 yildagi haqiqiy nashrning proportsional miqdoriga asoslangan edi.

Ushbu toifalar, kichik toifalar va namunalar soni to'g'risida qaror qabul qilingandan so'ng, haqiqiy namunalarni tanlash turli xil tasodifiy usullar bilan amalga oshirildi, birinchi navbatda, mavzu sohasida mavjud nashrlarning umumiy ro'yxatiga qo'llaniladigan tasodifiy raqamlar jadvalidan foydalanish. Namuna boshlanadigan sahifa ham tasodifiy raqamlar jadvali bilan tanlangan. Har bir namuna tanlangan sahifadagi birinchi to'liq jumladan boshlanadi. Sarlavhalar va bosh sarlavhalar, shuningdek, izohlar, jadvallar va rasm sarlavhalari olib tashlandi. 2000 ta so'zning

[4] Martin, Greer. Children in Progressive-Era America. Digital Public Library of America. September 2015.

taxminiy hisobi amalga oshirildi va keyingi jumlalar tanaffusida namuna tugatildi. Ushbu hisoblash maqsadlari uchun so'z oxirgi kodlashda (quyida tavsiflanishi kerak) har ikki tomonda bo'sh joy bo'lgan har qanday belgilar qatori sifatida aniqlandi, paragraf boshlanishining kod ko'rsatkichlari bundan mustasno. Gap bosh harf bilan boshlanib, yakuniy belgi (. ! yoki ?) bilan tugaydigan, keyin bo‘sh joy va bosh harf qo‘yiladigan, aniq qisqartmalar bundan mustasno. Ayrim hollarda gapning yakuniy belgisidan keyin bo‘sh joy qo‘yilmasligi mumkin; Keyinchalik aniqroq hisoblash kompyuter tomonidan amalga oshirildi.

Mualliflik huquqi bilan himoyalangan barcha materiallar uchun mualliflik huquqi egasining ruxsati olingan. Mualliflik huquqi ruxsati tafsilotlari 33-176-betlardagi namunalarning alohida ro'yxatiga kiritilgan.

Korpusning oltita versiyasi mavjud. Hammasi bir xil asosiy matnni o'z ichiga oladi, lekin ular tipografiya va formatda farqlanadi.

1. A shakli. Bu Korpusning asl shakli, chunki u 1963-64 yillarda tayyorlangan. O'sha paytda kompyuterda chop etish vositalarining cheklovlari uning quyida 3-bo'limda tasvirlangan batafsil kodlash protsedurasidan foydalanishni talab qilgan.

2. B shakli. Bu "chizilgan" versiya bo'lib, undan defis, apostrof va formulalar va ellips belgilaridan tashqari barcha tinish belgilari va kodlari olib tashlangan. Bu, ayniqsa, individual so'zlarga qiziquvchilar uchun foydalidir hamda Kuchera va Frensisdagi chastotalar jadvallarini tayyorlashda ishlatilgan, Hozirgi Amerika ingliz tilining hisoblash tahlili (Providence: Broo’n University Press, 1967).

3. C shakli. Bu qisman o'chirilgan matndan foydalaniladigan "yorliqli" versiya bo'lib, unda faqat tegishli ismning bosh harfi va grammatik ahamiyatga ega bo'lgan tinish belgilari saqlanib qolgan. Ushbu versiyadagi har biri alohida so'z (token) 81 ta ro'yxatdagi grammatik tegga ega bo'lib, ularning har biri ma'lum bir so'z sinfini belgilaydi.

4. Bergen Shakl I. Bu versiya va quyidagi Gumanistik tadqiqotlar Norvegiya hisoblash markazida (gumanistisk forskning uchun NAVF ning EDB-senter) doktor Jostein Hauge rahbarligida Bergen universitetida tayyorlangan. Ikkalasida ham katta va kichik harflar, oddiy tinish belgilari va minimal maxsus kodlar mavjud. Ushbu versiyada tipografik ma'lumotlar saqlanadi va satr oxiridagi so'zlar hech qachon bo'linmasligidan tashqari, asl nusxadagi kabi bir xil bo'linish ishlatiladi.

5. Bergen II shakli. Ushbu versiyada tipografik ma'lumotlar biroz qisqartiriladi va yangi uzunroq satr ishlatiladi. Ushbu versiya EDB-senterdan (Harald Haarfagresgt. 31, Bergen universiteti, N-5007 Bergen, Norvegiya) to'liq KO’IC muvofiqligi bilan birga mikrofixda mavjud.

6. Brown MARC shakli. Ushbu versiya Stenford universitetida tayyorlangan. U katta matnli korpuslar uchun mos bo'lgan ikkita tez-tez qo'llaniladigan tadqiqot usullariga mos keladigan tarzda ishlab chiqilgan:

1. Izlash mezoni sifatida bitta so'z yoki so'z + kontekstdan foydalangan holda to'liq jumlali iqtiboslarni qidirish va olish;

2. Kalit so'zning turli xil tartiblari va undan oldingi yoki keyingi og'zaki kontekstga ko'ra tashkil etilishi mumkin bo'lgan KO’IC shaklidagi muvofiqliklarni yaratish.

Xulosa va takliflar. Xulosa qilib aytganda, Brown korpusi va boshqa birinchi

avlod korpuslari hisoblash tilshunosligi va tabiiy tilni qayta ishlashning rivojlanishida muhim rol o`ynaydi. Ular tadqiqotchilarga ishlash uchun katta hajmdagi ma`lumotlarni taqdim etdilar va bugungi kunda ham qo`llanilayotgan ko`plab asosiy tushunchalar va usullarni yaratishga yordam berdilar.

Adabiyotlar:

1. Francis, W. Nelson & Henry Kucera. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press. 1967.

2.Biber, D. *Nutq Va Yozish Bo‘Yicha O‘Zgarishlar* . Kembrij: Kembrij Universiteti Nashriyoti, 1988 Yil.

3.Biber, D. Ko'p O'lchovli Yondashuvlar. In: Lüdeling, A.; Kytö, M. (Tahr.). *Korpus Lingvistikasi* - Xalqaro Qo'llanma. Berlin / Nyu-York: Valter De Gruyter, 2009 Yil.

4.Biber, D.; Conrad, S. *Registr, Janr Va Uslub* . Kembrij; Nyu-York: Kembrij Universiteti Nashriyoti, 2009. (Tilshunoslik Bo'yicha Kembrij Darsliklari).

5. Winthrop Nelson Francis and Henry Kučera. Frequency Analysis of English Usage: Lexicon and Grammar, Houghton Mifflin. 1983.

6.Cameron, L. *Ta'lim Nutqidagi Metafora* . London: Continuum, 2003.

7.McEnery, Tony & Wilson, Andrew. Corpus Linguistics. Edinburgh:Edinburgh University Press. 2001.

8.McEnery, Tony, Yukio Tono & Xiao, Richard. Corpus based Language Studies: An Advanced Resourse Book.London:Routledge. 2006.

9. Mukherjee,Joybrato. Anglistische Korpuslinguistik. Eine Einfuhrung. Berlin:Erich Schmidt. 2009.

10.Scherer, Carmen. Korpuslingvistik. Eine Einfuhrung. Heidelberg: Winter.2006.

11. Biber, Douglas et al. Corpus Linguistics:Investigating Language Structure and Use. Cambridge:CUP. 1998.