INTERNET MATNLARINING TIL KORPUSI RIVOJIDAGI AHAMIYATI

Abdullayeva Oqila Xolmo‘minovna
ToshDO‘TAU katta o‘qituvchisi, filologiya fanlari bo‘yicha falsafa doktori PhD
Norpulotova Munira Shomurod qizi
ToshDO‘TAU o‘qituvchisi

https://doi.org/10.5281/zenodo.7189101

Annotatsiya. *Korpus yaratishda internetdan foydalanish qulay. Chunki korpusda biror tilga oid yozma matnlar elektron variantda yoki kitob holatida bo‘lsa, skaner qilinib, vord (word) shakliga o‘tkaziladi, bunda esa matnda xatoliklar ko‘plab uchrab, to‘g‘rilanishi talab etiladi. Audio matnlarning esa transkripti yozilishi kerak. Mana shu jarayonda yozma matnlarning tayyor elektron variantidan til korpusini qurish afzal ko‘riladi. Veb tilning tabiati to‘g‘risida kutilmagan savollarni taqdim etishi mumkin. Shuningdek, u matn bilan ishlash va tekshirish uchun qulay vositani taqdim etadi.*

Kalit so‘zlar*: korpus, internet, vebkorpus, webCorp, NLP, BNC.*

ЗНАЧЕНИЕ ИНТЕРНЕТ-ТЕКСТОВ В РАЗВИТИИ ЯЗЫКОВОГО КОРПУСА

Аннотайия. *Для создания корпуса удобно использовать Интернет. Потому что, если письменные тексты определенного языка есть в электронном варианте или в виде книги, они сканируются и переводятся в словесную (словную) форму, и в тексте встречается много ошибок, которые нужно исправлять. Аудиотексты должны быть расшифрованы. В этом процессе предпочтительно строить языковой корпус из готовых электронных версий письменных текстов. Сеть может поставить неожиданные вопросы о природе языка. Он также предоставляет удобный инструмент для работы с текстом и его проверки.*

Ключевые слова: *корпус, интернет, вебкорпус, webCorp, NLP, BNC.*

THE IMPORTANCE OF INTERNET TEXTS IN THE DEVELOPMENT OF THE LANGUAGE CORPUS

Abstract. *It is convenient to use the Internet to create a corpus. Because if the written texts of a certain language are in the electronic version or in the form of a book, they are scanned and converted into word (word) form, and many errors in the text are encountered and need to be corrected. Audio texts should be transcribed. In this process, it is preferred to build a language corpus from the ready-made electronic version of written texts. The Web can present unexpected questions about the nature of language. It also provides a convenient tool for working with and checking text.*

Keywords: *corpus, internet, webcorpus, webCorp, NLP, BNC.*

KIRISH

Til ma’lumotlari kabi turli xil tillarga boy bo‘lgan internet juda ko‘p miqdorda va erkin ravishda mavjud bo‘lib, bu tilshunoslarning o‘yin maydonchasidir. Ba’zi tadqiqotchilar to‘g‘ridan-to‘g‘ri tijorat qidiruv tizimlaridan chastota ma’lumotlarini to‘plashadi. Boshqalar tegishli sahifalarni topish uchun qidiruv tizimidan foydalanadilar, korpus sifatida veb sahifalardagi matnli materiallarni tahlilga tortadi. Yana bir guruh mutaxassislar veb orqali korpus quradilar va keyin ular to‘plagan ma’lumotlarni boshqarish bilan shug‘ullanadilar. Xuddi shu tarzda, tilshunoslar uchun internet qidiruv tizimlarining ba’zi prototiplari taklif qilingan. Masalan, Webcorp

(http://www.webcorp.org.uk/) veb sahifasi so‘z birikmalarini yaratishga qodir. Korpus sifatida veb-saytlardan foydalanganda quyidagi imkoniyatlarga ega bo‘ladi:

- lug‘at, glossariy, tezaurusga kirish;
- ontologiyalarga kirish (masalan, WORNET);
- kollokatsiyalarni tahlil qilish;
- Google kabi qidiruv tizimi orqali iboralarni tahlil qilish;
- internetdagi yangiliklar hisobotining qiyosiy tahlili;
- parallel korpuslarni qurish (ko‘plab veb-sahifalar turli tillarga tarjima qilingan);
- paydo bo‘layotgan yangi leksik birliklarni o‘rganish (tilning yangi qo‘llanilishi);
- internetdagi ijtimoiy tarmoqlarni o‘rganish;
- ixtisoslashgan korpuslarni o‘rganish (akademik, biznes, yangiliklar va boshqa korpuslar);
- veb-janrlarni o‘rganish.

Internet korpus sifatida konsepti Kilgarriff va Grefenstettelar tomonidan qo‘llanilgan va internet korpusmi savolini o‘rtaga tashlashgan. Internet korpusmi yoki yo‘qligini aniqlash uchun, avvalo, korpus nima ekanligini aniqlashni taklif etishgan. Makkenri va Vilson prinsipial ravishda bir nechta matnlardan iborat har qanday to‘plamni korpus deb atash mumkinligini aytishgan. Ammo “korpus” atamasi zamonaviy tilshunoslik kontekstida ishlatilganda, ko‘pincha ushbu sodda ta’rifga qaraganda o‘ziga xos murakkab ma’nolarga ega.

TADQIQOT METODI VA METODOLOGIYASI

Makkenri va Hardielar veb korpus sifatida tushunchasini ko‘p jihatdan monitor korpusiga o‘xshatadi, chunki u har doim o‘sib boradigan katta hajmdagi ma’lumotlar to‘plami sifatida qurilgan va undan tilni o‘rganish uchun foydalanadi. Internetdan yoki vebdan korpus sifatida foydalanish uchun google kabi standart qidiruv tizimlaridan foydalanish bilan bir qatorda, tadqiqotchilar ushbu veb foydalanishni qo‘llab-quvvatlash uchun maxsus ishlab chiqilgan WebCorp (Renouf, 2003) interfeysini ham ishlab chiqdilar. WEbCorp veb-sayt hisoblanib, u lingvistik ma’lumotlarni olish uchun mo‘ljallangan. WebCorpda lingvistik tadqiqotlar uchun maxsus ishlab chiqilgan qidiruv variantlari mavjud (konkordans, so‘zlar ro‘yxati va boshqalar). WebCorp vebsayti lingvistik ma’lumotlarni olish uchun mo‘ljallangan: foydalanuvchining qidiruv natijasini qaysi kontekstda bo‘lganligini ko‘rsatadigan konkordans ro‘yxati mavjud. Ammo WebCorp vebsayti korpus yoki qidiruv tizimlariga nisbatan sekin ishlashi va ma’lum bir vaqt olishi bilan foydali hisoblanmaydi.

Lingvistik tadqiqotlar uchun korpus sifatida vebning o‘rganilishi 90-yillardan boshlangan deyish mumkin, sababi aynan o‘sha davrlarda internet matnlaridan korpusga materiallar yig‘ish sifatida foydalanish bir buncha o‘sdi. Radev va Makkoun tillarni yaratish tizimi uchun ma’lumot manbayi sifatida internetdagi yangiliklar tarmog‘idan foydalangan. So‘nggi yillarda barcha tadqiqotchilar internetda mavjud maqolalar to‘plamidan, xususan, qo‘llanmalar, dissertatsiyalar, ilmiy maqolalar yozishda yoki muayyan sohalar bo‘yicha natijaviy xulosalarga ega bo‘lishda foydalanishmoqda. Grefenstette, Nioke va Jons hamda Ganilar veb-resurslarning imkoniyatlarini elektron resurslar kam bo‘lgan tillar uchun til korpuslarining manbayi deya ta’kidlashgan, Resnik esa ikki tilli parallel korpuslar manbayi sifatida o‘rgangan. Fuji va Ishikava ensiklopediya yozuvlarini yaratish uchun internetdan foydalanishganini yozishgan. Grefenstette leksik ma’lumot manbayi sifatida internetga oid istiqbollar va tajribalarni taqdim etgan, chunki veb ko‘plab tillar uchun minglab kontekstli misollarni taqdim etadi, tillar uchun leksik yozuvlarni empirik dalillardan avtomatik ravishda topish imkoniyatlarini yaratadi. Bu kabi soha o‘zining yangiligi va

kirish xarajatlari yo‘qligi bilan talabalar va boshqa tadqiqotchilarni jalb qiladi. Umuman, yuqoridagi tadqiqotlar ro‘yxati to‘liq emas. Bu esa veb-korpus sifatida foydalanish qanday tez rivojlanayotganligini ko‘rsatadi.

TADQIQOT NATIJASI

Korpus tilshunosligi uchun veb-saytdan foydalanish juda yangi tendensiya. Kompyuter lingvistikasiga tegishli bo‘lgan yondashuvlar soni hali ham oz. Ammo allaqachon veb turli xil lingvistik darajadagi vazifalar uchun sinab ko‘rilgan. Masalan, leksikografiya, sintaksis, semantika va tarjima yo‘nalishida ko‘plab tadqiqotlar olib borilyapti. Volk aynan mana shu to‘rtta yo‘nalishda vebdan korpus sifatida yondashuvning o‘ziga xos jihatlarini bayon qilgan. Leksikografiya bo‘yicha eng muhim vazifalarni veb bajara oladi, chunki boshidanoq internet leksik resurslarni (turli tillardagi so‘zlar ro‘yxati) to‘plash va tarqatish uchun manba bo‘lgan. Ammo kompyuter nuqtayi nazaridan yanada qiziqarli jihati internetdagi matnlarning boyligidan yangi leksik materialni topish va tasniflashdan iborat. Bunga yangi so‘zlarni yoki iboralarni topish, tasniflash va odatdagi so‘z birikmalariga, subkategoriyalarga qo‘yiladigan talablar yoki ta’riflar kabi qo‘shimcha ma’lumotlarni to‘plash kiradi. Internetda to‘g‘ri nomlarni o‘rganish va tasniflash usulini tahlil qila oladi. Bu juda muhim ish, chunki to‘g‘ri nomlar doimiy ravishda yangi nomlar ixtiro qilinadigan ochiq so‘zlar sinfidir. Ularning tizimi uch bosqichda ishlaydi. Birinchidan, yig‘uvchi qidiruv tizimi tomonidan olingan vebsahifalarni so‘rovdan so‘ng kalit so‘z qolipiga yuklab oladi. Ikkinchidan, ro‘yxatlar, jadvallardan nomlarni topish uchun tahlilchilar ishlatiladi. Uchinchidan, filtrlash moduli nomlarni yetakchi aniqlovchilardan yoki o‘zaro bog‘liq bo‘lmagan so‘zlardan tozalaydi. Shu kabi qator vazifalarni sanoqli daqiqalar ichida veb bajaradi. Keyingi yo‘nalish, sintaksis bo‘yicha internetda ba’zi jumlalarni, hatto kichikroq matnlarni tahlil qiluvchi vebsahifalar mavjud. Lekin Volk fikricha, butunjahon internet tarmog‘i (WWW) qidiruv tizimlari lingvistik so‘rovlar uchun emas, balki umumiy bilim so‘rovlari uchun tayyorlanadi. Internetdan semantik ma’lumot to‘plash ham juda katta vazifa hisoblanadi, bu leksik birliklarni (nomlarni) qidirishga o‘xshaydi, ba’zan natija kutilganidan hayratlanarli bo‘lishi mumkin. Chunki bir nom qidirilganda, uning turli semantik ma’nolariga qo‘shimcha sinonimlari va boshqa tegishli so‘zlar ham hamroh bo‘ladi. Tarjima xizmati bo‘yicha internetdan foydalanish eng yuqori cho‘qqiga chiqqani hech kimga sir emas. Bir vaqtning o‘zi butun boshli matn yoki vebsahifalarning turli tillarda berilishi bu ham internetning eng muhim imkoniyatlaridan biridir.

MUHOKAMA

Ingliz-Braun va LOB uchun birinchi standart korpuslar yaratilganidan buyon tezkor texnologik o‘zgarishlar, nafaqat korpus analitik vositalarining tarqalishiga olib keldi, balki korpus hajmi jihatidan katta qadamlar tashlanishiga imkon berdi. Zamonaviy korpuslarning hajmi million emas, 100 million so‘zdan iborat bo‘lib, mutaxassislar turli xil materiallardan foydalanishni boshlagan. Bunga sabab qilib bir necha faktlar keltiriladi:

1. Korpus tilshunoslikning ba’zi sohalari uchun, hatto Britaniya milliy korpusi (BNC) tipidagi yangi mega o‘lchamdagi korpuslar ham unchalik katta emas. Leksik innovatsiyalarning morfologik samaradorligini o‘rganish haqiqatan ham yangi megakorpuslardan chetga chiqadigan materiallarga muhtoj bo‘lgan.
2. Texnologik rivojlanishlarning o‘zi yangi matn turlarini keltirib chiqardi, ulardan Braun korpuslari elektron pochtadan boshqasiga duch kelmagandi. Ularda chat xonasida muzokaralar, matnli xabarlar, bloglar yoki internet jurnallar mavjud edi, bu matn turlari qiziqarli o‘rganish obyekti hisoblanadi. Bundan tashqari matn turlariga yozma va og‘zaki so‘zlarni muhokama qilish

uchun yangi o‘lchovni qo‘shdi. Chunki ularning barchasi yozma vositadan foydalanadilar, ammo biz og‘zaki nutqda ko‘rishimiz mumkin bo‘lgan qoliplarga juda yaqin. Ba’zi tashqi variatsiyalar uchun an’anaviy matn turlari mavjud emas, ammo elektron pochta almashinuvida matnli xabar almashish, bloglar yoki interfaol xatlardan foydalanish mumkin. Va nihoyat, butun dunyo vebsaytidagi (www) yangi matn turlari, “xususiy til” singari ijtimoiy-pragmatik hodisalarni muhokama qilishda qiziqish uyg‘otadi.

3. Standart ma’lumotnoma korpuslarini tuzish uchun ko‘p vaqt va katta moliyaviy mablag‘lar talab etiladi, ular yaqin yillarda davom etayotgan o‘zgarishlarda tezda eskiradi.

4. Internetning o‘zida tildan foydalanish doimiy ravishda o‘zgarib turadigan tilni o‘zgartirish uchun asosiy manba bo‘lishi mumkin. “Weblish” yoki “netspeak”ning tilimizga ta’sirini baholash uchun biz ushbu hodisaning o‘zi haqida yaxshiroq tushunchaga ega bo‘lishimiz kerak.

Birinchi navbatda, korpusga asoslangan lingvistik tadqiqotlarda internetdan foydalanilgan 2 xil yo‘lni ajratib olishimiz kerak:

1) veb korpus sifatida tijorat brauzerlari yoki internetga asoslangan qidirish dasturlari yordamida, balki yanada tizimli tarzda bir evristik vosita sifatida veb korpusdan foydalanishi mumkin;

2) internet, shu bilan bir qatorda, katta oflayn monitor korpuslarini (veb korpus qurilishi uchun) tuzish uchun manba sifatida foydalanishi mumkin.

XULOSA

Dragomir Radev korpus lingvistikasi, umuman, tabiiy tilni qayta ishlash (NLP)da vebning ham muhim xizmatlari, imkoniyatlari borligini ta’kidlaydi. D.Radev tabiiy tilni qayta ishlash mexanizmida “berish” va “olish” o‘rtasida foydali farqni aniqladi. NLP vebsahifalar yoki veb qidiruv natijalarini mashina tarjimasi, ko‘p tilli hujjatlarni olish, savolga javob berish va nafaqat kerakli hujjatni, balki hujjatning kerakli qismini topish uchun boshqa strategiyalarni belgilash, tahlil qilish va boshqa asosiy texnologiyalarni umumlashtirib, veb texnologiyalar deb hukm qiladi. “Qabul qilish yoki olish” bu vebsaytni har qanday korpus lingvistikasi yoki NLP maqsadlari uchun ma’lumot manbai sifatida ishlatish sanaladi. Uzoq muddat davomida vebning lingvistik mohiyatini va, umuman olganda, boshqa maqsadlarda manbalar bilan to‘ldirib boramiz. Buning uchun biz vebsaytni o‘zini cheklangan holda o‘rganish obyekti sifatida qabul qilishimiz kerak. Ko‘pgina veb qidiruv texnologiyalari til texnologiyasiga asoslangan holda ishlab chiqilgan. Shuningdek, veb ko‘p tillarni qamrab olganligi bilan multilingvistik inklyuziv (kutilayotgan xizmatlar) hisoblanadi, shu bilan birga tanlash huquqiga ega eklektik vosita (eng yaxshisini tanlay olish imkoniyati) ham deyish mumkin. Aynan bu kabi imkoniyatlar internetning korpus sifatida baholanishiga sabab bo‘ldi.

REFERENCES

1. Kilgarriff A., Grefenstette G. Introduction to the special issue on the Web as corpus. Computational Linguistics, 29 (3), 2003. pp. 333-347.

2. Terra E., Clarke C. Frequency estimates for statistical word similarity measures. In Proceedings of the Human Language Technology and North American Chapter of Association of Computational Linguistics Conference 2003, 244-251.
3. Stuart K. New perspectives on corpus linguistics. file:///D:/about%20corpora/Dialnet-NewPerspectivesOnCorpusLinguistics-1426958.pdf
4. McEnery T., Wilson A. Corpus Linguistics. Edinburgh University Press, Edinburgh, 1996.
5. McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012.
6. http://www.webcorp.org.uk/
7. Radev D., McKeown K. Building a generation knowledge source using internet-accessible newswire. In proceedings of the Fifth Applied Natural Language Processing conference. Washington D. C., April 1997, pp. 221-228
8. Grefenstette G., Nioche J. Estimation of English and non-English Language Use on the WWW. In proceedings of RIAO (Recherche d'Informations Assistee par Ordinateur), Paris, 2000
9. Jones R and Ghani R. Automatically building a corpus for a minority language from the web. 38th Meeting of the ACL, Proceedings of the Student Research Workshop. Hong Kong. October 2000, pp. 29-36
10. Resnik P. Mining the web for bilingual text In proceedings of the 37th Meeting of ACL. Maryalnd, USA, June 1999, pp. 527-534.
11. Fujii A., Ishikawa T. Utilizing the world wide web as an encyclopaedia: Extracting term descriptions from semi-structured text. In proceedings of the 38th Meeting of the ACL, Hong Kong, October 2000, pp. 488-495
12. Grefenstette G. The WWW as a Resource for Example-Based MT Tasks. Invited Talk, ASLIB 'Translating and the Computer' conference, London. October, 1999.
13. Volk M. Using the Web as Corpus for linguistic research. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.6964&rep=rep1&type=pdf
14. Hundt M., Nesselhauf N. C.Biewer. Corpus linguistics and the web. – Amsterdam-New York, 2007. (https://books.google.co.uz/books?id=SdsA_xydxyQC&pg=PA69&dq=corpus+linguistics&hl=ru&sa=X&ved=0ahUKEwjco42UzqrkAhVBw4sKHSCACu84ChDoAQhVMAY#v=onepage&q=corpus%20linguistics&f=false)