



EXTRALINGUISTIC TAGGING OF PARALLEL CORPUS TEXTS OF UZBEK-ENGLISH PHRASES

Oydin Nurullayeva

Doctoral student of Alisher Navoi Tashkent State University of Uzbek

Language and Literature

<https://doi.org/10.5281/zenodo.13888531>

ARTICLE INFO

Received: 26th September 2024

Accepted: 29th September 2024

Online: 30th September 2024

KEYWORDS

Corpus linguistics, corpus of parallel texts, linguistic tagging, extralinguistic tagging, metadata, annotation.

ABSTRACT

In this article, extralinguistic tag or meta-information (Russian: метаданные) is a bibliographic, typological, thematic, sociological description covering external-intellectual information; the type of tag that combines the formal-structural tag (text, section, chapter, part, paragraph, sentence), as well as the technical-technological tag (encoding, processing date of the electronic version of the source), and its importance. Also, the options provided by the metadata set to the user are mentioned.

O'ZBEK-INGLIZ TILI FRAZEMALARINING PARALLEL KORPUSI MATNLARINI EKSTRALINGVISTIK TEGGLASH

Oydin Nurullayeva

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti tayanch
doktoranti

<https://doi.org/10.5281/zenodo.13888531>

ARTICLE INFO

Received: 26th September 2024

Accepted: 29th September 2024

Online: 30th September 2024

KEYWORDS

Korpus lingvistikasi, parallel matnlar korpusi, lingvistik teg, ekstralingvistik teg, metaaxborot, annotatsiya.

ABSTRACT

Mazkur maqolada ekstralingvistik teg yoki metaaxborot (ruscha: метаданные) – tashqi-intellektual ma'lumotlarni qamrab oluvchi, bibliografik, tipologik, tematik, sotsiologik tavsifni; formal-struktur teg (matn, bo'lim, bob, qism, xatboshi, gap)ni, shuningdek, texnik-texnologik teg (kodirovka, manbaning elektron versiyasini qayta ishlash sanasi)ni birlashtiruvchi teg turi haqida, uning ahamiyati haqida so'z boradi. Shuningdek, metaaxborot majmuining foydalanuvchiga beradigan imkoniyatlari keltirib o'tiladi.

Korpus lingvistikasida parallel matnlar korpusi ham muhim ahamiyat kasb etadi. Parallel matnlar korpusi badiiy asar, qo'llanma, ommaviy axborot vositalari, turli xil hujjatlarning ikki yoki undan ko'p tillardagi elektron ko'rinishlaridir. Parallel korpuslar yordamida bir so'z, jumla, abzats, supersintaktik butunliklarning turli tillardagi variantlarini



bilish mumkin. Parallel korpuslar madaniyatlararo muloqot keng yoyilgan hozirgi davr uchun muhim voqelik hisoblanadi.

Parallel matnlar korpusi avtomatik tarjima rivoji uchun ham xizmat qiladi, kompyuter leksikografiyasining taraqqiyotini ta'minlaydi. Parallel matnlar korpusi yordamida maxsus konkordanser dasturlar ishlab chiqiladi va turli xil ixtisoslik lug'atlari tuzish imkoniyati yuzaga keladi.

Badiiy asarlar asosidagi parallel matnlar korpusida frazemalar bilan bog'liq muammolar bo'lishi aniq. O'zbek tilidagi frazemalar ikki yoki undan ortiq so'zdan tuzilib, shu so'zlar semantikasi asosida yangi ko'chma lug'aviy ma'noni shakllantirishga xizmat qiladi. Shu bois parallel matnlar korpusini shakllantirishda to'g'ridan-to'g'ri mashina tarjimasidan foydalanilsa, semantikani to'g'ri yoritishda muammolar kelib chiqadi. Frazemalar tarjimasida, albatta, mutaxassis nazorati kerak bo'ladi. [3]

V.P. ta'rifi ko'ra. Zaxarov va S.Yu. Bogdanova: "Matnlarning lingvistik yoki til korpusi - bu ma'lum bir muammoni hal qilish uchun mo'ljallangan, birlashtirilgan, tuzilgan, etiketlangan, filologik jihatdan to'g'ri keladigan katta, mashina o'qilishi mumkin bo'lgan format. Lingvistik muammolar ... Zamonaviy korpusning asosiy xususiyatlari - mashinada o'qiladigan format, reprezentativlik va metalingvistik ma'lumotlarning mavjudligi" [4].

Turli lingvistik topshiriq / amalni bajarish uchun matnga lingvistik va ekstralingvistik qo'shimcha ma'lumot bilan ishlov berilgan bo'lishi lozim. Buning uchun mavjud matnning komponentlariga maxsus izoh (masalan, so'z turkumi haqida ma'lumot) berilishi zarur hisoblanadi. Bu izoh matn annotatsiyasi, maxsus lingvistik belgilar esa teg (belgi, ishora) (rus. razmetka, ing. tag) deb ataladi. Matn birliklarini izohlash esa annotatsiyalash yoki lingvistik izohlash (annotation), har bir lingvistik birlikni ramzlash esa teglash (разметка, tagging, mark- 346 up) deyiladi. Masalan, ot turkumi "Ot" yoki "N", sifat turkumi "Sif" yoki "Adj", fe'l turkumi "F" yoki "V" tarzida teglanadi. Matnda esa ot turkumiga tegishli "kitobim" so'zshakli "Ot, n-b nom., birl., I sh., b. eg.q." kabi annotatsiyalanadi, ya'ni lingvistik izohlanadi. Teglash va annotatsiyalashning eng oddiy misoli - so'z turkumlarini umumiy ramzlash va umumiy izohlash hisoblanadi.

Shuningdek, til korpuslarini ekstralingvistik teglash ham muhim ahamiyatga ega. Razmetka mutaxassislar tomonidan ikki: lingvistik va ekstralingvistik turga ajratiladi. Tadqiqotchi Sh.Hamroeva V.P.Zaxarov, S.Yu.Bogdanovalarning fikriga tayanib, ekstralingvistik razmetkaning quyidagi turlarini farqlaydi [2]:

1. Matn formatining o'ziga xosligini aks ettiruvchi (bob, xatboshi, qism va h.k.) teg.
2. Matn, uning muallifiga tegishli ma'lumotni ifodalovchi teg.

Chunki muallif haqidagi ma'lumot nafaqat nom, balki yosh, jins, u yashagan yil kabi ma'lumotlarni ham bildirishi mumkin. Matn haqidagi ma'lumot esa matn (asar) nomidan tashqari tili, yozilgan hamda nashr etilgan yilini qamrab oladi. Bunday ma'lumotlarning mavjudligi bazada detallashgan qidiruvni amalga oshirish imkonini beradi.

Ekstralingvistik teg yoki metaaxborot (ruscha: metadannye) - tashqi-intellektual ma'lumotlarni qamrab oluvchi, bibliografik, tipologik, tematik, sotsiologik tavsifni; formal-struktur teg (matn, bo'lim, bob, qism, xatboshi, gap)ni, shuningdek, texnik-texnologik teg (kodirovka, manbaning elektron versiyasini qayta ishlash sanasi)ni birlashtiruvchi teg turi hisoblanadi. Metaaxborot majmui foydalanuvchiga uning imkoniyatlarini aniqlab beradi. Bu



ma'lumotlarni tanlashda korpus tuzuvchisi tadqiqot maqsadi, tilshunoslar talabi hamda matnga u yoki bu qo'shimcha belgini qo'shish imkoniyati bilan tanish bo'lmog'i lozim. Tashqi intellektual teg, birinchidan, tilning o'zaro aloqasi, mavjudligini aniqlash; ikkinchidan, tilning o'ziga xos xususiyatini o'rganish uchun kerak. Tilga ta'sir qiladigan ikkita: tashqi, ya'ni nolisoniy hamda ichki omil mavjud. Dj. Sinkler tashqi omilning uch guruhini ajratadi:

1. Muallif tomonidan matn yaratishga aloqadorlik omili.
2. Matnning tashqi belgilariga tegishli omil.
3. Matnning yaratilishiga sabab bo'luvchi, auditoriyaga ta'sir qiluvchi omil.

Parallel korpuslarning ilk prototiplari faqatgina matnni abzaslar yoki gaplar kesimida tenglashtirishga qaratilgan bo'lsa, hozirgi kunda yaratilayotgan korpuslarning imkoniyatlari bundan ancha keng ko'lamda ekanligini kuzatishimiz mumkin. Quyida nisbatan oddiy darajada lingvistik va ekstralingvistik teglangan Aleuska(nemiz-istan-bask) uch tilli parallel korpusi bilan tanishamiz. Naroa Zubillaga, Zurine Sanz va Ibon Uribarrilar [4] tomonidan yaratilgan korpus uchta tilni o'z ichiga qamrab oladi. Korpusning maqsadi nemis tilidan bask tiliga qilingan bevosita tarjimalarni o'rganish maqsadida yaratilgan. Bask tiliga istan tilidan bilvosita qilingan tarjimalar ham o'rin olgan. Korpusni yaratish jarayonida olimlar Trace-Aligner dasturini ham yaratishgan bo'lib, ushbu dastur orqali matnlar abzas va gaplar kesimida tenglashtirilishi bilan birga, ular ekstralingvistik ma'lumotlar bilan ham boyitilganiga guvohbo'lishimiz mumkin.

Trace Aligner dasturi interfeysi. Dastlab dastur orqali ma'lumotlarga ekstralingvistik izoh beriladi: avval matnlar avtomatik XML formatda annotatsiyalanadi, "header" qismga matn haqidagi metatavsiflar beriladi; keyingi jarayonda gaplar avtomatik ravishda tenglashtiriladi; so'nggi bosqichda esa tenglashtirilgan avtomatik ma'lumotlarning to'g'rilik darajasi inson omili orqali tekshiriladi. Metatavsiflarni berish korpus uchun muhim omil deb sanalgan, ularning maqsadi korpus menejeri orqali so'rovlar bajarilganda foydalanuvchilarga qulaylik yaratish bo'lgan.

Parallel korpus tuzishda ma'lumotlar bazasining ahamiyati va uni tuzish prinsiplari ekstralingvistik razmetka bilan chambarchas bog'liq. Ma'lumki, parallel korpusning ahamiyati kontrastiv grammatikani o'rganish bilan belgilanadi.

Parallel korpus materiali sifatida matnni tanlashda, asosan, mukammal tarjimasi mavjud o'zbekcha asarlarni saralab olish tamoyili to'g'ri, degan fikrga keldik. Bunday matnlarga ekstralingvistik izoh yozishda o'zbek tilidagi asliyat va ingliz tilidagi tarjima matnlariga quyidagi izohlarni yozish maqsadga muvofiq bo'ladi.

Birinchidan, o'zbek tilidagi matn formatining bob, xatboshi, qism va h.k. qismlari ajratiladi, belgilanadi hamda ingliz tilidagi tarjima matnining aynan o'sha bob, xatboshi, qismlari belgilanib, ikkala til uchun matn qismlari tenglashtiriladi: asliyatda nechta qism bo'lsa, tarjima matnda qismlar soni shuncha bo'lishi kerak. Tarjimada biror xatboshi tushib qolgan taqdirda, bunday qismlar ko'rsatiladi. Biz o'zbek-ingliz parallel korpusi uchun matn tanlash jarayonida asar qismlari (xatboshi)ning tengligiga alohida e'tibor qaratamiz. Chunki parallel korpus foydalanuvchisi ikki tildagi matnda barcha qismlarni parallel kuzatib, o'rganib borishi lozim.

Ikkinchidan, matnning ekstralingvistik razmetkasiga tegishli keyingi izoh matn, uning muallifiga tegishli ma'lumotni ifodalovchi teg bo'lishi lozim. O'zbek-ingliz parallel korpusining



o'zbek tili qismida asar muallifi haqidagi ma'lumot uning nomi, yoshi, jinsi, yashagan yili kabi ma'lumotlardan iborat bo'lsa, ingliz tilidagi qismida shu ma'lumotlar ingliz tilida beriladi. Shu bilan birga asar tarjimoni haqida ham xuddi shu parametrlar asosida ma'lumotlar biriktiriladi.

Uchinchi, o'zbek va ingliz tili – ikkala qismda matn haqidagi ma'lumot matn (asar) nomi, tili (o'zbek/ingliz), yozilgan hamda nashr etilgan yili kabi izohdan iborat bo'ladi. Bunday ma'lumotlarning mavjudligi korpusning ikkala qismida detallashtirilgan qidiruvni birdek amalga oshirish imkonini beradi. Agar shunday izoh faqat aslyat matnida bo'lsa-yu, tarjima matnga ekstralingvistik izoh berilmagan bo'lsa, qidiruv teng natija bermaydi. Foydalanuvchi faqat bitta tildagi matnning ma'lumotlarini o'qiydi. Parallel korpusning eng asosiy xususiyati ikkala tildagi matnning barcha xususiyatlarini birdek kuzatish imkonini berish ekan, ekstralingvistik ma'lumotlar ham shunday biriktirilishi, matnlar shu prinsip asosida razmetkalanishi lozim. [1]

O'zbek-ingliz parallel korpusining ekstralingvistik razmetka majmui aslyat matni uchun matn muallifining ismi, familiyasi, taxallusi, jinsi, tug'ilgan vaqti, matn nomi, yozilish vaqti va joyi, matn hajmi: matnning necha so'zdan iboratligi haqidagi ma'lumotdan iborat bo'ladi. Tarjima matni uchun ushbu ma'lumotlarga matn tarjimoni, tarjima vaqti, joyi, matnning tarjimadagi hajmi haqidagi ma'lumot qo'shiladi. Parallel korpusning ikkala qismida ham matn janri, matn tipi, matn xronotopining ko'rsatilishi ekstralingvistik razmetkani mukammallashtiradi.

References:

1. Каримов Р. Ўзбек-инглиз параллел корпусини тузишнинг лингвистик ва дастурий масалалари. Филол. фан. Бўйича фалсафа докт. (PhD) диссертацияси. – Бухоро, 2022
2. Хамроева Ш. Ўзбек тили муаллифлик корпусининг тузишнинг лингвистик асослари: Филол. фан. Бўйича фалсафа докт. (PhD) диссертацияси. – Бухоро, 2018. – Б. 72-73.
3. Холбеков М. Ўзбекистонда таржима ва таржимашунослик (1991-2016 йиллар) // Хорижий филология. – 2016 йил. – №3. – Б. 16-42.
4. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. –Иркутск: ИГЛУ, 2011. – Б.36.