



Aziz YULDASHEV,
ToshDO'TAU o'qituvchisi
E-mail: yuldashevaziz@navoiy-uni.uz

F.f.d, professor Sh.Hamroyeva taqriza asosida

ICHKI KORPUS (PODKORPUS) ISHLAB CHIQISHDA EKSTRALINGVISTIK TEGTLARDAN FOYDALANISH

Аннотация

Ushbu maqolada ichki korpus (podkorpus) ishlab chiqishda ekstralingvistik teglardan foydalanish haqida fikr yuritiladi. Ichki korpus (podkorpus) – bu katta lingvistik korpusning muayyan mezonlar bo'yicha ajratilgan qismidir. Bunda milliy korpus turli xil ichki korpuslarni o'z ichiga oladi. Ichki korpus tilshunos va tadqiqotchilarga tilning muayyan jihatlariga e'tibor qaratish, uni batafsil o'rganish imkonini beradi.

Kalit so'zlar: Ichki korpus, podkorpus, ekstralingvistik teg, janr, mavzu, uslub.

ИСПОЛЬЗОВАНИЕ ЭКСТРАЛИНГВИСТИЧЕСКИХ ТЕГОВ ПРИ РАЗРАБОТКЕ ВНУТРЕННЕГО КОРПУСА (ПОДКОРПУСА)

Аннотация

В данной статье рассматривается использование экстралингвистических тегов при разработке внутреннего корпуса (подкорпуса). Внутренний корпус (подкорпус) — это часть более крупного лингвистического корпуса, которая выделяется в соответствии с определенными критериями. В данном случае национальный корпус включает в себя различные внутренние корпуса. Внутренний корпус позволяет лингвистам и исследователям сосредоточиться на конкретных аспектах языка и изучить его подробно. Использование внутренних корпусов позволяет проводить точный и целенаправленный анализ конкретных языковых данных.

Ключевые слова: Внутренний корпус, подкорпус, экстралингвистический тег, жанр, тема, стиль.

USING EXTRALINGUISTIC TAGS IN DEVELOPING AN INTERNAL CORPUS (SUBCORPUS)

Annotation

This paper discusses the use of extralinguistic tags in the development of an internal corpus (subcorpus). An internal corpus (subcorpus) is a part of a larger linguistic corpus that is identified according to certain criteria. In this case, a national corpus includes various internal corpora. An internal corpus allows linguists and researchers to focus on specific aspects of a language and study it in detail. The use of internal corpora allows for precise and targeted analysis of specific linguistic data.

Key words: Internal corpus, subcorpus, extralinguistic tag, genre, topic, style.

Kirish. Ichki korpus (podkorpus) – bu katta lingvistik korpusning muayyan mezonlar bo'yicha ajratilgan qismidir. Bunda milliy korpus turli xil ichki korpuslarni o'z ichiga oladi. Masalan:

Janr bo'yicha: Badiiy adabiyotlar korpusi, Ilmiy adabiyotlar korpusi, Publitsistik asarlar korpusi, So'zlashuv nutqi korpusi, She'riyat korpusi.

Yaratilish vaqti bo'yicha: XIX asr matnlari korpusi, XX asr matnlari korpusi, Zamonaviy matnlar korpusi.

Muallif bo'yicha: Muayyan yozuvchining asarlari korpusi.

Boshqa mezonlar bo'yicha: Gazeta korpusi, Og'zaki nutq korpusi.

Asosiy qism. Ichki korpuslarning ahamiyati. Ichki korpus tilshunos va tadqiqotchilarga tilning muayyan jihatlariga e'tibor qaratish, uni batafsil o'rganish imkonini beradi. Masalan, ichki korpus vositasida quyidagilarni tahlil qilish va o'rganish mumkin:

muayyan davr tilining lisoniy xususiyatlari;
janrlar o'rtasidagi uslubiy farqlar;
yozuvchining individual uslubi (yozuvchi idiolekti)
so'zlashuv nutqining xususiyatlari.

Ichki korpuslardan foydalanish ma'lum bir til ma'lumotlarini aniq va maqsadli tahlil qilish imkonini beradi.

Ichki korpus shakllantirish tor sohaga yo'naltirilgan tadqiqotlar olib borish uchun umumiy til korpusidan muayyan qismni tanlash jarayonidir. Kichik korpuslarni shakllantirishning asosiy usullari sifatida quyidagilar ajratiladi:

Mavzuviy mezon bo'yicha: Asosiy so'zlarni aniqlash bosqichida tadqiqotchini qiziqitirgan mavzuga aloqador so'zlar ro'yxati tuziladi. Korpusda qidiruvni amalga oshirish uchun shu so'zlardan foydalaniladi. Metama'lumotlardan foydalanish.

Korpus metama'lumotlarni (masalan, janr, muallif, nashr yili) o'z ichiga olsa, matnlar mavzu bo'yicha tanlanadi.

Janr mezon bo'yicha: Janr xususiyatlari bo'yicha tanlash. Tadqiqotchini qiziqitirgan janrga xos belgilar (masalan, badiiy adabiyot, ilmiy maqola, publitsistika). Aniqlanadi. Matnlarni tanlashda tadqiqot obyektini talash muhim ahamiyat kasb etadi. Metama'lumotlardan foydalanish. Korpus metama'lumotlarni o'z ichiga olsa, ulardan janr bo'yicha filtrlash funksiyasi o'rnatiladi va undan foydalaniladi.

Vaqt (davr) mezon bo'yicha: Nashr sanasi bo'yicha tanlash uchun vaqt chegaralari belgilanadi va shu davrda nashr etilgan matnlar tanlanadi. Metama'lumotlardan foydalanish. Korpus metama'lumotlarni o'z ichiga olsa, ulardan nashr yili bo'yicha filtrlash uchun foydalanish mumkin.

Mualliflik mezon bo'yicha: Mualliflik bo'yicha tanlash uchun muayyan muallif yoki mualliflar guruhining asarlari tanlanadi. Bunda ham metama'lumotlardan foydalanish talab etiladi.

Morfologik yoki sintaktik mezonlar bo'yicha: Teglar tazimidan foydalanish. Korpus morfologik yoki sintaktik xususiyatlar bo'yicha teglangan bo'lsa, matnlarni tanlashda ushbu teglashdan foydalanish lozim. Masalan, muayyan sintaktik konstruksiyaga ega bo'lgan barcha gaplarni tanlash mumkin[1]. Regulyar ifodalardan foydalanish. Muayyan morfologik yoki sintaktik belgiga ega bo'lgan so'z yoki iboralarni qidirish uchun regulyar ifodalardan foydalanish mumkin.

Ixtisoslashtirilgan dasturiy vositalardan foydalanish:

Korpus menejerlari. Ko'pgina korpus menejerlari[2] ichki korpuslarni shakllantirish uchun ishchi vositalarni taqdim etadi. Masalan, Rus tili milliy korpusi (NKRYA) foydalanuvchi kichik korpuslarini yaratish imkonini beradi.

Skript va dasturlar. Nisbatan murakkab vazifalar uchun Python yoki boshqa dasturlash tillarida skriptlardan foydalanish mumkin.

Ichki korpuslarni shakllantirishda e'tibor talab qilinadigan muhim jihatlar sifatida quyidagilarni aytish mumkin:

birinchidan, tanlash mezonlari aniq belgilanishi lozim;

ikkinchidan, korpus kerakli metama'lumotlarni o'z ichiga olganiga ishonch hosil qilish lozim, aks holda tanlangan parametrlar bo'yicha korpus shakllantirilmaydi.

uchinchidan, belgilangan vazifaga mos keladigan vositalardan foydalanish maqsadga muvofiq.

Milliy yoki umumiy korpus tarkibidagi ichki korpusni shakllantirishda ekstralingvistik teglar muhim ahamiyat kasb etadi, shuningdek, korpusning aniq maqsad va vazifalariga bog'liq. Podkorpusning shakllantirishda ekstralingvistik (nolisoniy) teglarning korpus bilan munosabatini quyida tavsiflaymiz:

1. Ichki korpus ekstralingvistik omillarni tahlil qilish uchun vosita sifatida quyidagi ahamiyatga ega. Ichki korpuslar ko'pincha ekstralingvistik mezonlar asosida shakllantiriladi[8], masalan: matn janri (badiiy adabiyot, ilmiy maqola, publitsistika); vaqt davri (XIX asr, zamonaviy nutq); mualliflik (muayyan yozuvchining asarlari); mavzu (iqtisod, tibbiyot, texnika). Demak, ichki korpusning o'zi ekstralingvistik omillarning tilga ta'sirini o'rganish uchun muhim vosita bo'lib xizmat qiladi.

2. Ichki korpusning tavsifi usuli sifatida ekstralingvistik teglar quyidagi ahamiyatga ega. Ekstralingvistik teglar korpusidagi matnlarni annotatsiyalash va ularning xususiyatlarini tavsiflash uchun ishlatiladi. Bu teglar quyidagilar haqida ma'lumotlarni o'z ichiga olishi mumkin: muallif; nashr sanasi; janr; mavzu. Ekstralingvistik teglar ichki korpus chegaralarini aniq belgilash, uning xususiyatlarini chuqur tahlil qilish imkonini beradi.

3. Ichki korpus va ekstralingvistik teglarning o'zaro ta'siri. Ichki korpus va ekstralingvistik teglar chambarchas bog'liq. Ichki korpus muayyan ekstralingvistik teglar asosida shakllantirilsa, ekstralingvistik teglar esa ichki korpus xususiyatlarini tavsiflashda ahamiyatli. Masalan, Milliy korpus ichida XIX asr badiiy adabiyoti ichki korpusini yaratish va uni "muallif", "janr", "mavzu" kabi ekstralingvistik teglar bilan annotatsiyalash mumkin.

4. Amaliy qo'llanilishi. Milliy korpuslar[9] tarkibidagi ichki korpuslar va ekstralingvistik teglar bilan ishlash uchun keng imkoniyatlarni taqdim etadi. Foydalanuvchi turli mezonlar asosida o'z ichki korpusini yaratishi, matn(lar)ni tahlil qilish uchun ekstralingvistik teglardan foydalanishi mumkin. Bu ekstralingvistik (nolisoniy) omillarning tilga ta'sirini hisobga olgan holda turli xil lingvistik tadqiqotlar olib borish imkonini beradi.

Xulosa qilib aytganda, ichki korpus va ekstralingvistik teglar bir-birini to'ldiruvchi vositalar bo'lib, ular lisoniy ma'lumotlarni chuqur va har tomonlama tahlil qilish imkonini beradi.

Milliy yoki umumiy korpusdan ichki korpusni shakllantirishning umumiy tamoyillari mavjud bo'lib, ularni quyidagi keltiramiz:

Filtrlash va faset qidiruv tamoyili. Foydalanuvchi ekstralingvistik ma'lumotlarga asoslangan filtrlardan foydalanib, o'z so'rovini aniqlashtirishi mumkin.

Ma'lumotlarni vizuallashtirish. Grafiklar, diagrammalar va boshqa vizual elementlar foydalanuvchiga kontekst va natijalar o'rtasidagi aloqalarni tushunishga yordam beradi.

Qidiruvni xususiyashtirish. Qidiruv tizimlari xususiy natijalarni taqdim etish uchun foydalanuvchining qidiruv tarixini hisobga olishi mumkin.

Xulosa sifatida aytish mumkinki, ekstralingvistik teglash qidiruv tizimini optimallashtiradi.

Sketch Engine[7]. Sketch Engine – matn korpusi bilan ishlash uchun onlayn vosita, u ichki korpuslarni yaratish imkonini ham beradi. Ichki korpuslarni tahlil qilish uchun bir qator vositalarni, jumladan, konkordans, statistik ma'lumotlar va vizuallashtirishni taqdim etadi. Interfeysi haqida. Sketch Engine foydalanuvchiga ichki korpuslarni oson yaratish va tahlil qilish imkonini beruvchi intuitiv interfeysga ega. Filtrlashning turli usullarini taqdim etadi. Masalan, tilshunos muayyan muallif uslubi(idiolekti)ni o'rganish uchun muallif tomonidan yozilgan barcha matnlarni qamrab olgan ichki korpus yaratishi mumkin. Kuzatishlarimiz shuni ko'rsatadiki, ichki korpus interfeysiga qo'yilgan talablar quyidagilardan iborat:

Moslashuvchanlik: foydalanuvchi turli mezonlar asosida ichki korpusni yaratish imkoniyatiga ega bo'lishi kerak;

foydalanishning qulayligi: interfeys intuitiv va foydalanish oson bo'lishi kerak;

ko'p funksiyalilik: interfeys ichki korpusni tahlil qilishda konkordans va statistik ma'lumotlar kabi vositalarni taqdim etishi kerak;

vizuallashtirish: ba'zi interfeyslar ma'lumotlarni vizuallashtirish vositalarini ham taqdim etadi, bu esa ichki korpusni tahlil qilishni osonlashtiradi.

Ichki korpus interfeysi lingvistik tadqiqot uchun muhim yordamchi vosita bo'lib, tadqiqotchiga tilning muayyan jihatiga e'tibor qaratish imkonini beradi. Katta korpusdan ichki korpus yaratish belgilangan mezonlar asosida korpusning muayyan qismini ajratish jarayonidir. Mazkur jarayonning asosiy bosqich va usullari quyidagicha tavsiflanadi[3]:

1-qadam: myezonlarni aniqlash. Ichki korpusga qaysi (qanday) matnlar kirishi kerakligi aniqlab olinadi. Bunda myezonlar turlicha bo'lishi mumkin: janr (badiiy adabiyot, publitsistika, ilmiy matn), muallif, nashr etilgan sana, til xususiyatlari (masalan, muayyan so'zlar yoki grammatik konstruksiyalardan foydalanish), mavzu, ekstralingvistik ma'lumotlar (muallifning jinsi, yoshi, yashash joyi) va h.

2-qadam: ichki korpus shakllantiruvchi vosita(lar)ni tanlash.

1. Biror bir milliy korpus (masalan, NKRYA) tanlanadi. NKRYA ichki korpus yaratish uchun mukammal qidiruv filtrlarini taqdim etadi[9]. Foydalanuvchi turli mezonlarni birlashtirib, murakkab so'rovni shakllantirishi mumkin (Qarang: 1-rasm).

The screenshot shows the 'Корпус' (Corpus) search interface. At the top, it displays statistics: '133 554 текста • 389 470 868 слов' (133,554 texts • 389,470,868 words). Below this, there are tabs for 'Основная информация' (Basic information) and 'Подкорпус' (Subcorpus). The 'Основная информация' tab is active, showing filters for 'Только тексты с автоматической разметкой' (Only texts with automatic markup) and 'Только тексты со своей структурой' (Only texts with their own structure). The 'Основные параметры текста' section includes fields for 'Название' (Name) set to 'Корпус Пушкина' (Pushkin Corpus), 'Автор' (Author) set to 'А.С. Пушкин', 'Дата создания' (Creation date), 'Язык' (Language) set to 'русский', and 'Год рождения' (Year of birth). The 'Жанр и тип текста' section has checkboxes for 'Нехудожественные тексты' (Non-fictional texts) and 'Художественные тексты' (Fictional texts), with a dropdown for 'Жанр текста' (Text genre) set to 'драматургия' (Drama).

1-rasm. Rus tili milliy korpusida ekstralingvistik filtrlar asosida ichki korpus shakllantirish jarayoni

2. Dasturiy usullar. Agar korpus fayllar ko'rinishida taqdim etilgan bo'lsa, matnlarni filtrlash uchun Python yoki boshqa dasturlash tillarida skriptlardan foydalanish mumkin. Belgilangan mezonlarga mos keladigan matnlarni qidirish uchun regulyar ifodalardan foydalanish mumkin.

3. Ma'lumotlar bazalari. Agar korpus ma'lumotlar bazasida saqlangan bo'lsa, ichki korpuslarni yaratish uchun SQL-so'rovlaridan foydalanish maqsadga muvofiq[1].

Korpus interfeysida ekstralingvistik teglarni ifodalash haqida. Korpus interfeysida ekstralingvistik teglarni ko'rsatish intuitiv va foydalanuvchi uchun qulay bo'lishi kerak. Buni amalga oshirishning bir necha usuli mavjud bo'lib, uning turlarini quyidagi aks ettiramiz.

Metama'lumotlar ko'rinishida: Jadval ko'rinishi: korpusdagi har bir matn uchun metama'lumotlar, jumladan, ekstralingvistik teglar jadval bilan ko'rsatiladi. Har bir teg alohida ustunda taqdim etiladi. Bu foydalanuvchiga turli matnlarning metama'lumotlarini tez ko'rib chiqish va qiyoslash imkonini beradi. Qalqib chiquvchi oynalar: kursor matn yoki uning nomi ustiga olib borilganda metama'lumotlar bilan qalqib chiquvchi oyna paydo bo'ladi. Bu interfeysni ortiqcha yuklamadan qutqaradi, tez ma'lumot olish uchun qulay.

Filtrlar va qidiruv ko'rinishida: Fasetli navigatsiya: foydalanuvchi matnlarni ekstralingvistik teglarning qiymatlari (masalan, janr, muallif, nashr etilgan sana) bo'yicha filtrlashi mumkin. Bu ekstralingvistik mezonlar asosida ichki korpus shakllantirish imkonini beradi. Teglar bo'yicha qidiruv: foydalanuvchi muayyan ekstralingvistik teglarning qiymatini o'z ichiga olgan matnlarni qidirishi mumkin. Masalan, muayyan

muallif tomonidan yozilgan yoki ma'lum bir mavzuga tegishli bo'lgan barcha matnlarni qidirish mumkin.

Vizualizatsiya ko'rinishida: Teglar buluti: ekstralingvistik teglar "teglar buluti" ko'rinishida ko'rsatiladi, bu yerda tegning o'lchami uning korpusdagi chastotasiga mos keladi. Bu foydalanuvchiga eng ommabop mavzu yoki mualliflarni tez ko'rish imkonini beradi. Grafik va diagramma: ekstralingvistik ma'lumotlar janr, muallif yoki nashr etilgan sanasi bo'yicha matnlarning taqsimlanishini ko'rsatuvchi grafik va diagramma ko'rinishida taqdim etilishi mumkin. Bu korpusdagi tendensiyalarni vizual tahlil qilishga yordam beradi.

Ma'lumotlarni batafsil ko'rsatish: Metama'lumotlar sahifasi: matn yoki uning metama'lumotlari ustiga bosilganda matn haqidagi batafsil ma'lumot, jumladan, barcha ekstralingvistik teglar bilan alohida sahifada ochiladi. Bu matn konteksti haqida to'liqroq ma'lumot olish imkonini beradi. Bog'langan matnlar: o'xshash ekstralingvistik teglarga ega bo'lgan matnlar ro'yxatini ko'rsatish mavzu jihatdan yaqin matnlarni o'rganish imkonini beradi.

Aytish joizki, interfeys turli darajadagi tayyorgarlikka ega bo'lgan foydalanuvchi uchun intuitiv va qulay bo'lishi, ma'lumotga tez kirish va qulay qidiruvni ta'minlashi, ma'lumotni vizualizatsiya qilish aniq va ma'lumotli bo'lishi kerak. Korpus interfeysida ekstralingvistik teglarni ko'rsatish foydalanuvchiga matnlarni chuqur tahlil qilish, ularning kontekstini hisobga olishga zamin yaratadi.

O'zbek tilida mavjud korpusdan "Zamonaviy o'zbek adabiyoti ichki korpusi"ni shakllantirish uchun ekstralingvistik omillarni hisobga olgan holda quyidagi teglardan foydalanish mumkin (Qarang: 1-jadval).

Tegning tipi	Teg mazmuni	Tegning qidiruvni shakllantirishdagi ahamiyati
Mavzuni ifodalovchi teglar	<Zamonaviy jamiyat> <Globalashuv> <An'analar va yangilik> <Ijtimoiy muammolar> <Tarix> <Shaxsiy munosabatlar> <Falsafiy mulohazalar>	Zamonaviy o'zbek jamiyati uchun dolzarb bo'lgan masalalar yoritilgan asarlarni ajratib ko'rsatish imkonini beradi.
Janriy xususiyatlarini ifodalovchi teglar	<Roman>, <Qissa>, <Hikoya>, <She'riyat>, <Drama>, <Esse>	Turli adabiy janrlarga mansub asarlarni ajratib ko'rsatadi.
Vaqt (asar yaratilgan yoki ifodalagan davr)ga oid teglar	Nashr yili: Aniq yil yoki vaqt davrini ko'rsatish (masalan, "2000-yillar")	Bu teglar ichki korpusni ma'lum bir davrda yaratilgan asarlar bilan cheklash imkonini beradi.
Muallif haqidagi teglar	<Muallif ismi, familiyasi>, <Muallif jinsi> <Muallif yoshi> <Muallifning yashash joyi>	Bu teglar aniq bir muallif yoki mualliflar guruhlarining asarlarini ajratib ko'rsatish imkonini beradi.
Ijtimoiy-madaniy teglar	<Shahar adabiyoti> <Qishloq adabiyoti> <Diaspora adabiyoti> <Ijtimoiy tarmoqlarning ta'siri>	Bu teglar zamonaviy o'zbek madaniyatining turli jihatlarini aks ettiruvchi asarlarni ajratib ko'rsatish imkonini beradi.

1-jadval. Zamonaviy o'zbek adabiyotiga oid asarlarning ekstralingvistik teglar tizimi mazmuni

Matnlarni ekstralingvistik teglashda ushbu teglardan foydalanish zamonaviy o'zbek adabiyotining xilma-xilligini aks ettiruvchi va uning ijtimoiy-madaniy kontekstini hisobga oluvchi ichki korpus yaratish imkonini beradi.

Xulosa. Xulosa sifatida aytish mumkinki, katta korpus ichidan ichki korpusni shakllantirishda ekstralingvistik teglarning mavzusi, mazmuni, ularni belgilash tamoyillari, qo'shimcha vositalarni to'g'ri tanlash muhim ahamiyat kasb etadi.

ADABIYOTLAR

- Christ O. A Modular and Flexible Architecture for an Integrated Corpus Query System Oliver Christ // Proceedings of COMPLEX'94 3rd Conference on Computational Lexicography and Text Research Budapest, Hungary, July 7-10, 1994, pp. 23-32.
- https://en.wikipedia.org/wiki/Corpus_manager
- <https://guides.libraries.psu.edu/c.php?g=1139753&p=8391284>
- The Uppsala Russian Corpus: <http://www.slaviska.uu.se/korpus.htm> (05.12.2024); The Lancaster-Oslo/Bergen Corpus of British English, LOB: <http://clu.uni.no/icame/manuals/LOB/INDEX.HTM> (accessed: 04.12.2024);
- <https://www.english-corpora.org/bnc/>; Brown Corpus:
- <https://uzschoolcorpara.uz/>
- <https://www.sketchengine.eu/>
- Manios K, Nenadic G, Spasic I, Ananiadou S. An Integrated Term-Based Corpus Query System. In {EACL} 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 12-17, 2003, Agro Hotel, Budapest, Hungary. - 2003. p. 243-250.
- <https://ruscorpora.ru/>