O‘ZBEKISTON RESPUBLIKASI

OLIY TA’LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

TOSHKENT DAVLAT SHARQSHUNOSLIK UNIVERSITETI

“TASDIQLAYMAN”
G.Sh.Rixsiyeva
2025-yil “ 20 ” 06

TIL KORPUSLARI
FANINING O‘QUV DASTURI

Bilim sohasi:	200 000 - San’at va gumanitar fanlar
Ta’lim sohasi:	230 000 – Tillar
Ta’lim yo‘nalishi:	60230400 Kompyuter lingvistikasi

Toshkent – 2025

Sharq xalqlari tillari va adabiyoti
instituti direktori: _____ X.V.Mirzaxmedova
(imzo)

Tarjimashunoslik, tilshunoslik va
xalqaro jurnalistika oliy maktabi boshlig'i: _____ S.T.Mustafayeva
(imzo)

Arm boshlig'i: _____ M.Yuldasheva
(imzo)

Fan/modul kodi	O‘quv yili	Semestr	YeCTS - Kreditlar
TK 16-06	2027 /2028	6	6

Fan/modul turi	Ta’lim tili	Haftadagi dars soatlari
Majburiy	O‘zbek	6

1.	Fanning nomi	Auditoriya mashg‘ulotlari (soat)	Ma’ruza	Amaliy	Mustaqil ta’lim (soat)	Jami yuklama (soat)
	Til korpuslari	72	36	36	108	180

2.

I. Fanning mazmuni:

Fanning maqsadi: Ushbu fan til korpuslari va ularning lingvistik tahlilda ahamiyatini chuqur o‘rganish, korpus yaratish, annotatsiya va analiz qilish usullarini o‘rgatishni maqsad qiladi. Talabalarga zamonaviy tilshunoslik va kompyuter lingvistikasi uchun muhim bo‘lgan korpus texnologiyalari va ularning qo‘llanilish sohalarini tushuntirish orqali til ma’lumotlarini tahlil qilish va o‘rganish imkoniyatlari taqdim etiladi.

Fan vazifasi - Til korpuslarining asosiy turlarini o‘rganish va ularning qo‘llanilishi sohalarini tushuntirish. Korpus lingvistikasi va tahlil usullari bo‘yicha asosiy tushunchalarni berish.Talabalarga til korpuslari bilan ishlash, ularni yaratish va annotatsiya qilish ko‘nikmalarini o‘rgatish. Lingvistik tahlil jarayonida korpuslardan foydalangan holda amaliy tajriba oshirish. Zamonaviy dasturiy vositalar orqali korpus ma’lumotlarini tahlil qilish, natijalarni interpretatsiya qilish va ulardan foydalanishni o‘rgatish.

Mazkur fan dasturi xalqaro tan olingan reytinglarda birinchi top 300 talik ro‘yxatga kiruvchi Lyudvig Maksimilian nomidagi Myunxen universiteti (Ludwig Maximilian University of Munich 44 ARWU) xorijiy tajribasini inobatga olgan holda takomillashtirildi.

https://www.cis.lmu.de/bsc/studienfach/studienfach/index.html

3.

II. Asosiy nazariy qism (ma'ruza mashg'ulotlari)

II.I. Fan tarkibiga quyidagi mavzular kiradi:

1-mavzu. Til korpuslariga kirish va til korpuslarining turlari

Til korpuslari – bu tilni o'rganish, tahlil qilish va rivojlantirish uchun tuzilgan yozma va og'zaki matnlar to'plamidir. Ushbu kursda talabalar korpuslarning tuzilishi, ularning tahlil usullari va tilshunoslikdagi ahamiyati bilan tanishadilar. Korpuslar orqali tilning haqiqiy foydalanishi, grammatik va leksik xususiyatlari, shuningdek, zamonaviy til o'zgarishlari o'rganiladi. Talabalar turli korpuslardan foydalanib, tadqiqotlar olib borish va tilni amaliy o'rganish ko'nikmalarini rivojlantiradilar. Turli xil matnlarni o'z ichiga oladi va tilning umumiy qoidalarini o'rganishga imkon beradi. Misol uchun, jurnal maqolalari, kitoblar va internet materiallari.Muayyan soha yoki mavzuga oid matnlarni o'z ichiga oladi, masalan, tibbiyot, biznes yoki huquq. Bu korpuslar soha tilini tahlil qilishda foydalidir. Turli mintaqalar va guruhlar orasidagi dialekt va lahjalarni o'rganish uchun mo'ljallangan. Ushbu korpuslar mintaqaviy farqlarni ko'rsatadi. Sifatli korpuslar muayyan matnlar va ularning kontekstini o'rganishga, miqdoriy korpuslar esa katta hajmdagi ma'lumotlarni tahlil qilishga mo'ljallangan. Talabalar ushbu turli korpuslarni tahlil qilib, ularning qo'llanilish sohalarini va til o'rganishdagi ahamiyatini o'rganadilar. Bu bilimlar lingvistika va tilshunoslikda muhim o'rin egallaydi.

2- mavzu. Korpuslar yaratishning bosqichlari va ma'lumot yig'ish usullari

Ushbu mavzuda talabalarga til korpuslarini yaratish jarayonining asosiy bosqichlari haqida ma'lumot beriladi. Korpus yaratish jarayoni bir necha muhim bosqichlardan iborat bo'lib, har bir bosqichning o'ziga xos ahamiyati bor.

Ma'lumot yig'ish, ma'lumotlarni tozalash va normalizatsiya qilish, annotatsiya va markup, korpus tuzilish, korpusni tahlil qilish. Ushbu mavzu talabalarga til korpuslarini yaratish jarayonini to'liq tushunish va amaliy ko'nikmalarni egallash imkonini beradi.

Ma'lumot yig'ish — korpus yaratish jarayonining muhim bosqichidir va uning to'g'ri amalga oshirilishi keyingi tahlil va tadqiqot natijalariga ta'sir qiladi. Manbalarni tanlash - Talabalar turli manbalardan ma'lumot yig'ish uchun tegishli manbalarni tanlashni o'rganadilar. Avtomatlashtirilgan yig'ish- Talabalar avtomatlashtirilgan ma'lumot yig'ish usullarini o'rganadilar, masalan, web-scraping va API-lardan foydalanish. Bu usullar yordamida katta hajmdagi ma'lumotlarni tez va samarali ravishda yig'ish mumkin. Anketalar va so'rovnomalar-Talabalar anketalar va so'rovnomalarni yaratish va ularni ma'lumot yig'ish jarayonida qanday qo'llashni o'rganadilar. Bu usul yordamida maqsadli guruhdan ma'lumot to'plash mumkin. Ushbu mavzu talabalarga ma'lumot yig'ish jarayonining asosiy usullarini tushunishga va amaliy ko'nikmalarni egallashga yordam beradi. Bu, shuningdek, korpus yaratishda sifatli ma'lumotlar to'plash uchun zarur bo'lgan metodologik bilimlarni shakllantiradi.

3- mavzu. Korpus tuzilishi va formatlari.Korpusni tozalash va normalizatsiya qilish

Ushbu mavzuda talabalarga til korpuslarining tuzilishi va saqlash formatlari haqida ma'lumot beriladi. Korpus tuzilishi uning qanday tashkil etilganini va ma'lumotlarning qanday shaklda saqlanishini belgilaydi. Korpusning to'g'ri tuzilishi va formatlari tadqiqot jarayonida ma'lumotlardan samarali foydalanishni ta'minlaydi. Korpus tuzilishi -Talabalar korpusning tuzilishini, ya'ni ma'lumotlarni qanday tartibga solish va tashkil etishni o'rganadilar. Korpus formatlari: Korpuslarni saqlash va almashish uchun foydalaniladigan

turli formatlar haqida ma'lumot beriladi. XML (eXtensible Markup Language). JSON (JavaScript Object Notation), CSV (Comma-Separated Values), YAML (YAML Ain't Markup Language). Ushbu mavzu talabalarga korpus tuzilishi va formatlarini to'g'ri tushunishga, ularni yaratish va ishlatish jarayonida amaliy ko'nikmalarni egallashga yordam beradi. Korpusni tozalash va normalizatsiya qilish — ma'lumotlarni tahlil qilishdan oldin ularning sifatini oshirish va tahlilga tayyorlash uchun muhim bosqichlardir.

Korpusni tozalash: Talabalar korpusdan shovqinlarni (noise) olib tashlash jarayoni bilan tanishadilar. Bu jarayon quyidagilarni o'z ichiga oladi. Ortqacha va kerakmas ma'lumotlarni olib tashlash, maxsus belgilarni va HTML teglarini olib tashlash, kichik harflarga o'tkazish, normalizatsiya qilish, lemmalash, stemming, morfologik normalizatsiya.

Ushbu mavzu talabalarga korpusni tozalash va normalizatsiya qilish jarayonining asosiy usullarini tushunishga va amaliy ko'nikmalarni egallashga yordam beradi.

4- mavzu. Annotatsiya va markup va korpusda morfologik belgilash

Ushbu mavzuda talabalarga annotatsiya va markup jarayonlari, ularning ahamiyati va korpuslar bilan ishlashdagi qo'llanilishi haqida ma'lumot beriladi. Annotatsiya va markup, ma'lumotlarni qayta ishlash va tahlil qilishda muhim rol o'ynaydi, chunki ular ma'lumotlarga qo'shimcha kontekst va tushuncha qo'shadi.

Annotatsiya: Talabalar annotatsiya jarayonini va uning maqsadlarini o'rganadilar. Annotatsiya — bu ma'lumotlar yoki matn segmentlariga qo'shimcha ma'lumotlar, belgilashlar yoki izohlar kiritish jarayonidir. **Markup:** Markup — bu matnning ma'lum qismiga maxsus belgilashlarni qo'shish jarayonidir. Talabalar markup formatlarini va ularning korpuslarda qanday ishlatilishini o'rganadilar

Ushbu mavzu talabalarga annotatsiya va markup jarayonlarining asosiy tamoyillarini tushunishga, ularni til korpuslarini yaratishda va ma'lumotlarni tahlil qilishda qanday qo'llashni o'rganishga yordam beradi. Morfologik belgilash — bu so'zlarning morfologik strukturasini tahlil qilish va ularning grammatik xususiyatlarini belgilash jarayonidir. Ushbu jarayon korpusni to'g'ri tushunish va tahlil qilish uchun zaruriy qadamdir. **Morfologik belgilashning maqsadi:** Talabalar morfologik belgilashning asosiy maqsadlari bilan tanishadilar. **Morfologik belgilash jarayonlari:** Talabalar morfologik belgilash jarayonining turli bosqichlari bilan tanishadilar. **Morfologik belgilash vositalari:** Ushbu mavzuda talabalarga morfologik belgilash uchun ishlatiladigan vositalar va algoritmlar haqida ma'lumot beriladi. Ushbu mavzu talabalarga korpusda morfologik belgilash jarayonining asosiy prinsiplarini tushunishga va amaliy ko'nikmalarni egallashga yordam beradi.

5- mavzu. Sintaktik annotatsiya va Semantik belgilash

Ushbu mavzuda talabalarga sintaktik annotatsiya jarayoni haqida ma'lumot beriladi. Sintaktik annotatsiya — bu til korpusidagi matnlarni sintaktik struktura va grammatik belgilari bilan belgilash jarayonidir. Ushbu jarayon yordamida matnning grammatik tuzilishi va uning tarkibidagi so'zlar o'rtasidagi munosabatlar aniqlanadi.

Semantik belgilash — bu matnda so'z va iboralarning ma'nosini aniqlash va ularning o'zaro kontekstual bog'liqligini belgilash jarayonidir. Bu jarayon yordamida matnning ma'no qatlamlari, tushunchalari va ularning o'zaro munosabatlari aniqlanadi. **Semantik rol belgilash:** Talabalar so'z va iboralarning jumladagi rolini belgilashni o'rganadilar. Bu bosqichda so'zlarning o'zaro bog'liqligi, masalan, subyekt, obyekt yoki vosita kabi rollar

aniqlanadi. **Entitetlarni aniqlash:** Talabalar matndan muhim tushunchalar va entitetlarni, masalan, shaxs, joy, vaqt yoki boshqa turdagi ob'ektlarni ajratishni o'rganadilar. Bu qadam matn mazmunini chuqurroq tushunishga yordam beradi. **Qo'llanma va vositalar:** Semantik belgilash uchun foydalaniladigan turli vositalar va texnikalar bilan tanishiladi. Talabalar WordNet, FrameNet kabi lug'atlar va qo'llanmalarni, shuningdek, semantik belgilash uchun dasturiy vositalarni o'rganadilar.Ushbu mavzu talabalarga matnlarni semantik nuqtai nazardan belgilashni va ularning ma'nosini aniqlashni o'rgatadi, bu esa tabiiy tilni qayta ishlashda chuqur tahlil va muloqot uchun muhim qadam hisoblanadi.

6- mavzu. Paralel korpuslar Til korpusida tez-tez uchraydigan iboralar

Ushbu mavzuda talabalarga parallel korpus tushunchasi va uning tabiiy tilni qayta ishlashda ahamiyati haqida ma'lumot beriladi. Parallel korpus — bir xil matnning turli tillarga tarjima qilingan versiyalarini o'z ichiga olgan matnlar to'plamidir. Bu korpuslar asosan tarjima jarayonlari va mashina tarjimasi uchun asosiy manba hisoblanadi. Parallel korpusning tuzilishi, qo'llanilishi, parallel korpuslar yaratish. Ushbu mavzu talabalar uchun til korpuslarining tarjima va ko'p tilli NLP jarayonlaridagi ahamiyatini tushunish imkoniyatini beradi. Tez-tez uchraydigan iboralar, yoki frazeologizmlar, ko'p hollarda tilda o'ziga xos ma'noni anglatadigan takroriy birikmalardan tashkil topadi. Ushbu iboralarni aniqlash va tahlil qilish turli tilshunoslik, tarjima va NLP jarayonlarida muhim ahamiyatga ega.Tez-tez uchraydigan iboralar va ularning xususiyatlari, aniqlash usullari, N-gram tahlili, statistik metodlar, qo'llaniladigan dasturiy vositalar qo'llanishi.Mazkur mavzu talabalar uchun til korpuslarida keng qo'llaniladigan iboralarni aniqlash va tahlil qilishda amaliy ko'nikmalarni egallash imkonini beradi.

7- mavzu. Morfologik va sintaktik analizatorlar bilan ishlash

Ushbu mavzuda talabalarga morfologik va sintaktik analizatorlarning qanday ishlashi va ularni turli tilshunoslik va tabiiy tilni qayta ishlash (NLP) jarayonlarida qanday qo'llash mumkinligi haqida umumiy tushuncha beriladi. Morfologik analizatorlar so'zlarning tuzilishini, ya'ni ularning leksik va grammatik qismlarini tahlil qilsa, sintaktik analizatorlar jumlalarning sintaktik tuzilishini va so'zlar o'rtasidagi munosabatlarni aniqlash uchun mo'ljallangan. Morfologik analizatorlar. Sintaktik analizatorlar. Qo'llanilish usullari va amaliy vositalar

8- mavzu. Korpuslardan kontekst olish va foydalanish

Bu mavzu talabalarga til korpuslarida so'z va iboralar uchun kontekstni olish va uni tahlil qilish usullarini o'rgatadi. Kontekstni olish — bu so'zlar yoki iboralar qanday atrof-muhitda, qaysi qo'shimchalar va grammatik shakllarda ishlatilishini aniqlash jarayoni. Ushbu usul so'z ma'nosini chuqurroq tushunish, tarjima sifatini oshirish, va til modellari qurishda muhimdir. Kontekst tushunchasi va uning ahamiyati, kontekstni olish usullari. Amaliy qo'llanilishi. Ushbu mavzu tilni chuqurroq anglash va tilda yashiringan mantiqiy bog'lanishlarni kashf etish uchun muhim bo'lgan kontekstni to'g'ri olish va undan samarali foydalanish bo'yicha bilim va ko'nikmalar beradi.

9- mavzu. Keyingi so'zlarni prognoz qilish

Keyingi so'zlarni prognoz qilish - bu sun'iy intellekt va tabiiy tilni qayta ishlash (NLP) sohasidagi texnologiya bo'lib, matnni davom ettirish uchun keyingi so'z yoki iborani taxmin qilish imkonini beradi. Ushbu texnologiya odatda til modellari yordamida amalga oshiriladi va chatbotlar, avtomatik matn yozish, va tarjima tizimlarida qo'llaniladi.

Studentlar bu mavzuda til modellarini qurish, ularning ishlash prinsiplari, ko‘p chastotali va n-gramm usullarini o‘rganadilar. Bu texnologiya so‘zlar orasidagi kontekstni tushunish va matnlarni aniq prognoz qilish uchun qanday ishlatilishini o‘rganishga yordam beradi.

10- mavzu. Til o‘rganishda korpuslardan foydalanish

Til o‘rganishda korpuslardan foydalanish - bu tilning lug‘at tarkibi, grammatik qoidalari va uslub xususiyatlarini real matnlar asosida o‘rganishni anglatadi. Korpuslar til o‘rganuvchilar uchun haqiqiy yozma va og‘zaki kontekstlar orqali o‘rganilayotgan tilning tabiiy ishlatilishini kuzatishga yordam beradi. Ular, ayniqsa, yangi so‘zlarni, iboralarni va ularning kontekstini o‘rganishda, grammatik va stilistik xatolarni kamaytirishda samarali hisoblanadi. Talabakar til o‘rganishda korpuslardan qanday foydalanishni, xususan, yangi so‘zlarni o‘rganish, grammatik qoidalarni tushunish va til me'yorlariga mos keladigan iboralarni topishda korpus qidiruv texnologiyalaridan foydalanishni o‘rganadilar. Shuningdek, korpuslarga asoslangan mashqlar yordamida so‘z boyligini oshirish, talaffuzni yaxshilash va til qoidalarini chuqurroq o‘rganish imkoniyatiga ega bo‘ladilar.

11- mavzu. Korpuslardan mashina o‘qitish ma’lumotlarini olish va ularni tozalash hamda tayyorlash

Korpuslardan mashina o‘qitish uchun ma’lumot olish - bu katta hajmdagi tabiiy til matnlaridan modelni o‘rgatish uchun zarur bo‘lgan ma’lumotlarni ajratib olish jarayonidir. Mashina o‘qitish modellari uchun korpuslar real hayotdagi til qoidalari, grammatik strukturalar va lug‘at tarkibiy qismlarini o‘zida mujassam etgani sababli juda muhim. Bu jarayonda korpuslardan so‘z chastotasi, ibora naqshlari, sintaktik va semantik ma’lumotlar kabi elementlarni to‘plash mumkin. Talabalar korpuslardan mashina o‘qitish uchun kerakli xususiyatlarni ajratish, tozalash, hamda modelni o‘rgatish uchun tayyorlash jarayonlarini o‘rganadilar. Ular, shuningdek, so‘z vektorlari, n-gramm usullari, va mashina o‘qitish uchun matnni xususiyatlarga aylantirish texnikalari haqida bilimga ega bo‘ladilar. Bu ko‘nikmalar tabiiy tilni qayta ishlashda samarali modellar yaratishga yordam beradi.

12- mavzu. O‘zbek tilida korpus yaratish tajribalari va ularni tahlil qilish hamda boshqarish

O‘zbek tilida korpus yaratish - bu tilshunoslik va tabiiy tilni qayta ishlash sohasida yangi imkoniyatlarni ochib beruvchi murakkab jarayon. O‘zbek tilidagi korpuslar cheklangan bo‘lgani uchun, ularni yaratish va rivojlantirishga katta ehtiyoj bor. Bu jarayonda matnlarni yig‘ish, tahrirlash va strukturalash asosiy bosqichlar hisoblanadi. Hozirgacha O‘zbek tilida korpus yaratishga qaratilgan bir qancha ilmiy loyihalar amalga oshirilgan, ular O‘zbek tilining lug‘at tarkibi, grammatikasi va sintaktik tuzilishini o‘rganishga xizmat qiladi.

Studentlar ushbu mavzuda O‘zbek tilida korpus yaratishning asosiy qiyinchiliklari va usullarini o‘rganadilar. Jumladan, matn to‘plash jarayonlari, yozuv tizimining o‘ziga xosliklari, dialektlar va uslubiy xususiyatlar kabi masalalar haqida bilimga ega bo‘ladilar. Shuningdek, korpusni tozalash, belgilarni normallashtirish, morfologik va sintaktik belgilash jarayonlari bilan tanishadilar. Bu bilimlar O‘zbek tilida mustahkam va funksional korpus yaratish imkonini beradi.

13- mavzu. Korpuslardan mashina tarjimasi uchun foydalanish

Korpuslardan mashina tarjimasi uchun foydalanish - bu tabiiy tilni qayta ishlash (NLP) sohasidagi muhim amaliyot bo‘lib, u ikki yoki undan ortiq til o‘rtasidagi tarjimalarni avtomatlashtirish uchun zarur bo‘lgan ma’lumotlarni taqdim etadi. Mashina tarjimasi uchun parallel korpuslar (bir xil mazmundagi matnlarning ikki yoki undan ortiq tildagi nusxalari)

ayniqsa foydalidir, chunki ular so‘z va iboralarning bir tilidan ikkinchisiga qanday tarjima qilinishini o‘rganish imkonini beradi. Bu usul, xususan, neyron tarmoqlarga asoslangan mashina tarjima tizimlarida keng qo‘llanadi.

Studentlar ushbu mavzuda parallel korpuslar bilan ishlash, ularni to‘plash va tahrirlash, tarjima sifatini oshirish uchun modelni o‘rgatish jarayonlarini o‘rganadilar. Shuningdek, segmentatsiya, so‘z birikmalarining ekvivalentlarini topish, grammatik moslik va kontekstual ma’nolarni saqlash kabi muhim texnikalar haqida bilimga ega bo‘ladilar. Bu ko‘nikmalar mashina tarjima tizimlarini yanada aniqroq va tabiiyroq qilishga xizmat qiladi.

14- mavzu. Til korpuslarida sozlamalar va izlash imkoniyatlari

Til korpuslarida sozlamalar va izlash imkoniyatlari - bu foydalanuvchilarga katta hajmdagi matnlar orasida kerakli ma’lumotlarni tez va samarali topishga yordam beruvchi funksiyalardir. Korpuslarda murakkab qidiruv mexanizmlari mavjud bo‘lib, ular orqali so‘zlar, iboralar, morfologik shakllar, sintaktik naqshlar, va kontekstual kombinatsiyalarni izlash mumkin.

Studentlar ushbu mavzuda korpuslar qidiruv tizimlarini, ular yordamida so‘z chastotasi, kontekstual qo‘llanilishi va grammatika naqshlarini qanday topish mumkinligini o‘rganadilar. Shuningdek, ma’lum sozlamalar orqali ma’lumotlarni filtr qilish, ya’ni maxsus kategoriyalar (masalan, mavzuga, uslubga yoki lahjaga qarab) bo‘yicha qidirish usullari bilan tanishadilar. Bu imkoniyatlar til tahlili, lingvistik tadqiqotlar va NLP dasturlarini yaratishda katta yordam beradi.

15- mavzu. So‘z birikmalarini tahlil qilish usullari va til korpuslarida variantli so‘z shakllari

So‘z birikmalarini tahlil qilish usullari - bu matnlarda tez-tez birga ishlatiladigan so‘zlar kombinatsiyalarini o‘rganish uchun qo‘llaniladigan usullardir. So‘z birikmalarining tahlili tilning leksik va sintaktik tuzilishini tushunishga yordam beradi, shuningdek, u tabiiy tilni qayta ishlash (NLP), mashina tarjimasi, va ma’nolarni aniqlash tizimlarida muhim o‘rin tutadi. Bu usullar yordamida so‘zlar orasidagi bog‘lanishlar, kontekstual ma’nolar va ibora naqshlari o‘rganiladi.

Studentlar ushbu mavzuda n-gramm (ikki so‘zli, uch so‘zli va undan ko‘p so‘zli kombinatsiyalar), chastotaviy tahlil, kollokatsiya, va o‘zaro axborot (mutual information) kabi asosiy usullarni o‘rganadilar. Shuningdek, ular korpuslardan foydalangan holda so‘z birikmalarini tahlil qilish, kombinatsiyalarning semantik xususiyatlarini aniqlash va kontekstual qo‘llanishiga qarab tahlil qilishni o‘zlashtiradilar. Bu ko‘nikmalar tilning tabiiy tuzilishini tushunishga va yanada samarali lingvistik tizimlar yaratishga imkon beradi.

Til korpuslarida variantli so‘z shakllarini aniqlash va tahlil qilish, tilning boyligini, ijtimoiy va kontekstual farqlarni tushunishga yordam beradi. Bunday shakllar til o‘rganishda, so‘z birikmalarini tahlil qilishda va grammatika tuzilmalarini o‘rganishda muhim ahamiyatga ega.

Talabalar bu mavzuda variantli so‘z shakllarining tahlilini, ularning kontekstdagi o‘zgarishini va qaysi faktorlar ularni shakllanishiga ta’sir qilishini o‘rganadilar. Ushbu tahlil, shuningdek, morfologik qayta ishlash va so‘zlarni normalizatsiya qilishda yordam beradi. Korpuslar yordamida variantli shakllarni tahlil qilish, talabalarni tilning nozik nuanslarini, sinonimlar va antonimlar o‘rtasidagi farqlarni va ularning ko‘plab kontekstlarda qanday ishlatilishini aniqlashga yo‘naltiradi. Bu jarayonlar, tilshunoslik, NLP va boshqa lingvistik tadqiqotlar uchun muhimdir.

16- mavzu. Matnni qayta ishlashda til korpuslari Til korpuslarining huquqiy va etik muammolari

Matnni qayta ishlashda til korpuslari - bu tabiiy tilni qayta ishlash (NLP) jarayonlarida foydalanuvchilarga haqiqiy matnlar asosida ko'plab til qoidalari va strukturalarini o'rganishga imkon beruvchi resurslardir. Korpuslar orqali turli xil tillarda yozilgan matnlar to'plamlarini taqdim etish, ularni tahlil qilish va modellar yaratish mumkin. Bu jarayonda korpuslar, avtomatik tarjima, matn tasnifi, sentiment tahlili, so'z birikmalarini aniqlash va boshqa ko'plab ilovalarda asosiy o'rin tutadi.

Studentlar ushbu mavzuda matnni qayta ishlash jarayonida korpuslardan qanday foydalanishni, ularni qanday to'plash, belgilash va tahlil qilishni o'rganadilar. Shuningdek, ular korpuslardan foydalanish orqali morfologik, sintaktik va semantik tahlil, jumlalarni segmentatsiya qilish va til modellari yaratish kabi vazifalarni qanday bajarish mumkinligini o'zlashtiradilar. Bu ko'nikmalar NLP sohasidagi amaliyot va tadqiqotlar uchun muhimdir, chunki ular yuqori sifatli til tizimlarini yaratishga yordam beradi.

Til korpuslarini yaratish va ulardan foydalanish jarayonida bir qator huquqiy va etik muammolar paydo bo'lishi mumkin. Ushbu muammolarni aniqlash va hal etish tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohasida muhimdir.

1. **Mualliflik huquqlari**: Korpuslarda ishlatiladigan matnlar ko'pincha mualliflik huquqiga ega bo'lgan materiallardir. Korpuslarni yaratishda yoki ulardan foydalanishda mualliflik huquqlarini buzmaslik uchun tegishli ruxsatnomalar olish zarur.
2. **Maxfiylik va shaxsiy ma'lumotlar**: Korpuslar, ayniqsa, ijtimoiy tarmoqlardan olingan ma'lumotlar, shaxsiy ma'lumotlarni o'z ichiga olishi mumkin. Shaxsiy ma'lumotlar va maxfiylikni himoya qilish muhim ahamiyatga ega bo'lib, foydalanuvchilar ma'lumotlarning qayerdan olinganini va qanday ishlatilishini bilishlari kerak.
3. **Nohaqlik va kamsitish**: Korpuslar turli ijtimoiy guruhlar, gender, yoki etnik identifikatsiyalarga nisbatan nohaq yoki kamsituvchi til namoyon etishi mumkin. Korpuslarni tahlil qilishda va modellarni yaratishda bunday muammolarga e'tibor berish zarur.
4. **Ommaviy foydalanish va qiyinchiliklar**: Korpuslardan foydalanishning qanday shartlari borligi haqida aniq va ochiq ma'lumot berish, ularning qayerda va qanday maqsadda foydalanilishini belgilash muhimdir.

Studentlar ushbu huquqiy va etik muammolarni o'rganish orqali til korpuslarini yaratish va ulardan foydalanishda zarur bo'lgan mas'uliyatli yondashuvni tushunadilar. Bunday bilimlar ularni etika va qonun doirasida ishlashga tayyorlaydi, bu esa lingvistik tadqiqotlar va NLP ilovalari uchun muhimdir.

17- mavzu. Mashhur til korpuslari va ularning xususiyatlari Brown Corpus

Til korpuslari, tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohalarida keng qo'llaniladigan resurslardir. Ularning bir nechtasi mashhur bo'lib, har biri o'ziga xos xususiyatlarga ega. Quyida ba'zi mashhur til korpuslari va ularning xususiyatlari keltirilgan: Brown Corpus, Penn Treebank, Corpus of Contemporary American English (COCA), ruscorpora.

Ushbu mashhur til korpuslari, talabalar va tadqiqotchilar uchun turli xil tilshunoslik va NLP vazifalarini bajarishda muhim resurs bo'lib xizmat qiladi. Har bir korpusning o'ziga xos

xususiyatlari, ulardan qanday foydalanish va tilni tahlil qilish jarayonida qanday qo‘llanilishi haqida bilish, tilshunoslik va informatika sohasidagi ko‘nikmalarni rivojlantirishga yordam beradi. Brown Corpus - bu tabiiy tilni qayta ishlash va tilshunoslik tadqiqotlarida keng qo‘llaniladigan eng mashhur korpuslardan biridir. 1960-yillarda (an'anaga ko'ra 1961 yilda) R. Hudzon va uning hamkasblari tomonidan yaratilgan. U ingliz tilidagi 1 million so‘zni o‘z ichiga oladi va turli janrlarda yozilgan matnlarni taqdim etadi. Brown Corpus, ingliz tilidagi eng muhim til resurslaridan biri bo‘lib, tilshunoslik, leksikologiya, stilistika va tabiiy tilni qayta ishlash sohalarida qimmatli manba hisoblanadi. U talabalarga va tadqiqotchilarga tilni chuqurroq o‘rganishga va yangi tadqiqotlar o'tkazishga yordam beradi.

18- mavzu. Til korpuslarini tahlil qilish uchun dasturiy vositalar Til korpuslarni yaratish

Til korpuslarini tahlil qilish jarayonida foydalaniladigan dasturiy vositalar tilshunoslar va tadqiqotchilarga matnlarni yanada samarali va aniq tahlil qilish imkonini beradi. Quyida mashhur dasturiy vositalar va ularning asosiy xususiyatlari keltirilgan: NLTK (Natural Language Toolkit), spaCy, Gensim, Sketch Engine, AntConc, Treetagger, extRazor. Ushbu dasturiy vositalar, talabalar va tadqiqotchilarga til korpuslarini tahlil qilishda yordam beradi, bu esa ularga tilning tuzilishi, leksik xususiyatlari va grammatik qoidalarini yanada chuqurroq tushunishga yordam beradi. Talabalarni til korpuslarini yaratish jarayonlari, ularning ahamiyati va foydalanish usullari bilan tanishtirish, shuningdek, tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohalarida korpuslardan qanday foydalanishni o‘rganishga yo‘naltirish. Ushbu mavzu orqali talabalar til korpuslarini yaratish jarayonini to‘liq tushunishga va kelajakda o‘z tadqiqotlarida ushbu resurslardan samarali foydalanishga tayyorlanadilar.

II.2. Amaliy mashg‘ulotlari bo‘yicha ko‘rsatma va tavsiyalar

“Til korpuslari ” fani bo‘yicha amaliy mashg‘ulotlar uchun quyidagi mavzular tavsiya etiladi:

1-mavzu. Til korpuslariga kirish va til korpuslarining turlari

Til korpuslari – bu tilni o‘rganish, tahlil qilish va rivojlantirish uchun tuzilgan yozma va og‘zaki matnlar to‘plamidir. Ushbu kursda talabalar korpuslarning tuzilishi, ularning tahlil usullari va tilshunoslikdagi ahamiyati bilan tanishadilar. Korpuslar orqali tilning haqiqiy foydalanishi, grammatik va leksik xususiyatlari, shuningdek, zamonaviy til o‘zgarishlari o‘rganiladi. Talabalar turli korpuslardan foydalanib, tadqiqotlar olib borish va tilni amaliy o‘rganish ko‘nikmalarini rivojlantiradilar. Turli xil matnlarni o‘z ichiga oladi va tilning umumiy qoidalarini o‘rganishga imkon beradi. Misol uchun, jurnal maqolalari, kitoblar va internet materiallari.Muayyan soha yoki mavzuga oid matnlarni o‘z ichiga oladi, masalan, tibbiyot, biznes yoki huquq. Bu korpuslar soha tilini tahlil qilishda foydalidir. Turli mintaqalar va guruhlar orasidagi dialekt va lahjalarni o‘rganish uchun mo‘ljallangan. Ushbu korpuslar mintaqaviy farqlarni ko‘rsatadi. Sifatli korpuslar muayyan matnlar va ularning kontekstini o‘rganishga, miqdoriy korpuslar esa katta hajmdagi ma'lumotlarni tahlil qilishga mo‘ljallangan. Talabalar ushbu turli korpuslarni tahlil qilib, ularning qo‘llanilish sohalarini

va til o‘rganishdagi ahamiyatini o‘rganadilar. Bu bilimlar lingvistika va tilshunoslikda muhim o‘rin egallaydi.

2- mavzu. Korpuslar yaratishning bosqichlari va ma’lumot yig‘ish usullari

Ushbu mavzuda talabalarga til korpuslarini yaratish jarayonining asosiy bosqichlari haqida ma’lumot beriladi. Korpus yaratish jarayoni bir necha muhim bosqichlardan iborat bo‘lib, har bir bosqichning o‘ziga xos ahamiyati bor.

Ma’lumot yig‘ish, ma’lumotlarni tozalash va normalizatsiya qilish, annotatsiya va markup, korpus tuzilish, korpusni tahlil qilish. Ushbu mavzu talabalarga til korpuslarini yaratish jarayonini to‘liq tushunish va amaliy ko‘nikmalarni egallash imkonini beradi.

Ma’lumot yig‘ish — korpus yaratish jarayonining muhim bosqichidir va uning to‘g‘ri amalga oshirilishi keyingi tahlil va tadqiqot natijalariga ta’sir qiladi. Manbalarni tanlash - Talabalar turli manbalardan ma’lumot yig‘ish uchun tegishli manbalarni tanlashni o‘rganadilar. Avtomatlashtirilgan yig‘ish- Talabalar avtomatlashtirilgan ma’lumot yig‘ish usullarini o‘rganadilar, masalan, web-scraping va API-lardan foydalanish. Bu usullar yordamida katta hajmdagi ma’lumotlarni tez va samarali ravishda yig‘ish mumkin. Anketalar va so‘rovnomalar-Talabalar anketalar va so‘rovnomalarni yaratish va ularni ma’lumot yig‘ish jarayonida qanday qo’llashni o‘rganadilar. Bu usul yordamida maqsadli guruhdan ma’lumot to‘plash mumkin. Ushbu mavzu talabalarga ma’lumot yig‘ish jarayonining asosiy usullarini tushunishga va amaliy ko‘nikmalarni egallashga yordam beradi. Bu, shuningdek, korpus yaratishda sifatli ma’lumotlar to‘plash uchun zarur bo‘lgan metodologik bilimlarni shakllantiradi.

3- mavzu. Korpus tuzilishi va formatlari.Korpusni tozalash va normalizatsiya qilish

Ushbu mavzuda talabalarga til korpuslarining tuzilishi va saqlash formatlari haqida ma’lumot beriladi. Korpus tuzilishi uning qanday tashkil etilganini va ma’lumotlarning qanday shaklda saqlanishini belgilaydi. Korpusning to‘g‘ri tuzilishi va formatlari tadqiqot jarayonida ma’lumotlardan samarali foydalanishni ta’minlaydi. Korpus tuzilishi -Talabalar korpusning tuzilishini, ya’ni ma’lumotlarni qanday tartibga solish va tashkil etishni o‘rganadilar. Korpus formatlari: Korpuslarni saqlash va almashish uchun foydalaniladigan turli formatlar haqida ma’lumot beriladi. XML (eXtensible Markup Language). JSON (JavaScript Object Notation), CSV (Comma-Separated Values), YAML (YAML Ain't Markup Language). Ushbu mavzu talabalarga korpus tuzilishi va formatlarini to‘g‘ri tushunishga, ularni yaratish va ishlatish jarayonida amaliy ko‘nikmalarni egallashga yordam beradi. Korpusni tozalash va normalizatsiya qilish — ma’lumotlarni tahlil qilishdan oldin ularning sifatini oshirish va tahlilga tayyorlash uchun muhim bosqichlardir.

Korpusni tozalash: Talabalar korpusdan shovqinlarni (noise) olib tashlash jarayoni bilan tanishadilar. Bu jarayon quyidagilarni o‘z ichiga oladi. Ortqacha va kerakmas ma’lumotlarni olib tashlash, maxsus belgilarni va HTML teglarini olib tashlash, kichik harflarga o‘tkazish, normalizatsiya qilish, lemmalash, stemming, morfologik normalizatsiya.

Ushbu mavzu talabalarga korpusni tozalash va normalizatsiya qilish jarayonining asosiy usullarini tushunishga va amaliy ko‘nikmalarni egallashga yordam beradi.

4- mavzu. Annotatsiya va markup va korpusda morfologik belgilash

Ushbu mavzuda talabalarga annotatsiya va markup jarayonlari, ularning ahamiyati va korpuslar bilan ishlashdagi qo‘llanilishi haqida ma’lumot beriladi. Annotatsiya va markup,

ma'lumotlarni qayta ishlash va tahlil qilishda muhim rol o'ynaydi, chunki ular ma'lumotlarga qo'shimcha kontekst va tushuncha qo'shadi.

Annotatsiya: Talabalar annotatsiya jarayonini va uning maqsadlarini o'rganadilar. Annotatsiya — bu ma'lumotlar yoki matn segmentlariga qo'shimcha ma'lumotlar, belgilashlar yoki izohlar kiritish jarayonidir. **Markup:** Markup — bu matnning ma'lum qismiga maxsus belgilashlarni qo'shish jarayonidir. Talabalar markup formatlarini va ularning korpuslarda qanday ishlatilishini o'rganadilar

Ushbu mavzu talabalarga annotatsiya va markup jarayonlarining asosiy tamoyillarini tushunishga, ularni til korpuslarini yaratishda va ma'lumotlarni tahlil qilishda qanday qo'llashni o'rganishga yordam beradi. Morfologik belgilash — bu so'zlarning morfologik strukturasini tahlil qilish va ularning grammatik xususiyatlarini belgilash jarayonidir. Ushbu jarayon korpusni to'g'ri tushunish va tahlil qilish uchun zaruriy qadamdir. **Morfologik belgilashning maqsadi:** Talabalar morfologik belgilashning asosiy maqsadlari bilan tanishadilar. **Morfologik belgilash jarayonlari:** Talabalar morfologik belgilash jarayonining turli bosqichlari bilan tanishadilar. **Morfologik belgilash vositalari:** Ushbu mavzuda talabalarga morfologik belgilash uchun ishlatiladigan vositalar va algoritmlar haqida ma'lumot beriladi. Ushbu mavzu talabalarga korpusda morfologik belgilash jarayonining asosiy prinsiplarini tushunishga va amaliy ko'nikmalarni egallashga yordam beradi.

5- mavzu. Sintaktik annotatsiya va Semantik belgilash

Ushbu mavzuda talabalarga sintaktik annotatsiya jarayoni haqida ma'lumot beriladi. Sintaktik annotatsiya — bu til korpusidagi matnlarni sintaktik struktura va grammatik belgilari bilan belgilash jarayonidir. Ushbu jarayon yordamida matnning grammatik tuzilishi va uning tarkibidagi so'zlar o'rtasidagi munosabatlar aniqlanadi.

Semantik belgilash — bu matnda so'z va iboralarning ma'nosini aniqlash va ularning o'zaro kontekstual bog'liqligini belgilash jarayonidir. Bu jarayon yordamida matnning ma'no qatlamlari, tushunchalari va ularning o'zaro munosabatlari aniqlanadi. **Semantik rol belgilash:** Talabalar so'z va iboralarning jumladagi rolini belgilashni o'rganadilar. Bu bosqichda so'zlarning o'zaro bog'liqligi, masalan, subyekt, obyekt yoki vosita kabi rollar aniqlanadi. **Entitetlarni aniqlash:** Talabalar matndan muhim tushunchalar va entitetlarni, masalan, shaxs, joy, vaqt yoki boshqa turdagi ob'ektlarni ajratishni o'rganadilar. Bu qadam matn mazmunini chuqurroq tushunishga yordam beradi. **Qo'llanma va vositalar:** Semantik belgilash uchun foydalaniladigan turli vositalar va texnikalar bilan tanishiladi. Talabalar WordNet, FrameNet kabi lug'atlar va qo'llanmalarni, shuningdek, semantik belgilash uchun dasturiy vositalarni o'rganadilar.Ushbu mavzu talabalarga matnlarni semantik nuqtai nazardan belgilashni va ularning ma'nosini aniqlashni o'rgatadi, bu esa tabiiy tilni qayta ishlashda chuqur tahlil va muloqot uchun muhim qadam hisoblanadi.

6- mavzu. Paralel korpuslar Til korpusida tez-tez uchraydigan iboralar

Ushbu mavzuda talabalarga parallel korpus tushunchasi va uning tabiiy tilni qayta ishlashda ahamiyati haqida ma'lumot beriladi. Parallel korpus — bir xil matnning turli tillarga tarjima qilingan versiyalarini o'z ichiga olgan matnlar to'plamidir. Bu korpuslar asosan tarjima jarayonlari va mashina tarjimasi uchun asosiy manba hisoblanadi. Parallel korpusning tuzilishi, qo'llanilishi, parallel korpuslar yaratish. Ushbu mavzu talabalar uchun til korpuslarining tarjima va ko'p tilli NLP jarayonlaridagi ahamiyatini tushunish imkoniyatini beradi. Tez-tez uchraydigan iboralar, yoki frazeologizmlar, ko'p hollarda tilda o'ziga xos

ma'noni anglatadigan takroriy birikmalardan tashkil topadi. Ushbu iboralarni aniqlash va tahlil qilish turli tilshunoslik, tarjima va NLP jarayonlarida muhim ahamiyatga ega.Tez-tez uchraydigan iboralar va ularning xususiyatlari, aniqlash usullari, N-gram tahlili, statistik metodlar, qo'llaniladigan dasturiy vositalar qo'llanishi.Mazkur mavzu talabalar uchun til korpuslarida keng qo'llaniladigan iboralarni aniqlash va tahlil qilishda amaliy ko'nikmalarni egallash imkonini beradi.

7- mavzu. Morfologik va sintaktik analizatorlar bilan ishlash

Ushbu mavzuda talabalarga morfologik va sintaktik analizatorlarning qanday ishlashi va ularni turli tilshunoslik va tabiiy tilni qayta ishlash (NLP) jarayonlarida qanday qo'llash mumkinligi haqida umumiy tushuncha beriladi. Morfologik analizatorlar so'zlarning tuzilishini, ya'ni ularning leksik va grammatik qismlarini tahlil qilsa, sintaktik analizatorlar jumlalarning sintaktik tuzilishini va so'zlar o'rtasidagi munosabatlarni aniqlash uchun mo'ljallangan. Morfologik analizatorlar. Sintaktik analizatorlar. Qo'llanilish usullari va amaliy vositalar

8- mavzu. Korpuslardan kontekst olish va foydalanish

Bu mavzu talabalarga til korpuslarida so'z va iboralar uchun kontekstni olish va uni tahlil qilish usullarini o'rgatadi. Kontekstni olish — bu so'zlar yoki iboralar qanday atrof-muhitda, qaysi qo'shimchalar va grammatik shakllarda ishlatilishini aniqlash jarayoni. Ushbu usul so'z ma'nosini chuqurroq tushunish, tarjima sifatini oshirish, va til modellari qurishda muhimdir. Kontekst tushunchasi va uning ahamiyati, kontekstni olish usullari. Amaliy qo'llanilishi. Ushbu mavzu tilni chuqurroq anglash va tilda yashiringan mantiqiy bog'lanishlarni kashf etish uchun muhim bo'lgan kontekstni to'g'ri olish va undan samarali foydalanish bo'yicha bilim va ko'nikmalar beradi.

9- mavzu. Keyingi so'zlarni prognoz qilish va matnni avtomatik tugatish

Keyingi so'zlarni prognoz qilish - bu sun'iy intellekt va tabiiy tilni qayta ishlash (NLP) sohasidagi texnologiya bo'lib, matnni davom ettirish uchun keyingi so'z yoki iborani taxmin qilish imkonini beradi. Ushbu texnologiya odatda til modellari yordamida amalga oshiriladi va chatbotlar, avtomatik matn yozish, va tarjima tizimlarida qo'llaniladi.

Studentlar bu mavzuda til modellarini qurish, ularning ishlash prinsiplari, ko'p chastotali va n-gramm usullarini o'rganadilar. Bu texnologiya so'zlar orasidagi kontekstni tushunish va matnlarni aniq prognoz qilish uchun qanday ishlatilishini o'rganishga yordam beradi.

10- mavzu. Til o'rganishda korpuslardan foydalanish

Til o'rganishda korpuslardan foydalanish - bu tilning lug'at tarkibi, grammatik qoidalari va uslub xususiyatlarini real matnlar asosida o'rganishni anglatadi. Korpuslar til o'rganuvchilar uchun haqiqiy yozma va og'zaki kontekstlar orqali o'rganilayotgan tilning tabiiy ishlatilishini kuzatishga yordam beradi. Ular, ayniqsa, yangi so'zlarni, iboralarni va ularning kontekstini o'rganishda, grammatik va stilistik xatolarni kamaytirishda samarali hisoblanadi.

Studentlar til o'rganishda korpuslardan qanday foydalanishni, xususan, yangi so'zlarni o'rganish, grammatik qoidalarni tushunish va til me'yorlariga mos keladigan iboralarni topishda korpus qidiruv texnologiyalaridan foydalanishni o'rganadilar. Shuningdek, korpuslarga asoslangan mashqlar yordamida so'z boyligini oshirish, talaffuzni yaxshilash va til qoidalarini chuqurroq o'rganish imkoniyatiga ega bo'ladilar.

11- mavzu. Korpuslardan mashina o‘qitish ma’lumotlarini olish va ularni tozalash hamda tayyorlash

Korpuslardan mashina o‘qitish uchun ma’lumot olish - bu katta hajmdagi tabiiy til matnlaridan modelni o‘rgatish uchun zarur bo‘lgan ma’lumotlarni ajratib olish jarayonidir. Mashina o‘qitish modellari uchun korpuslar real hayotdagi til qoidalari, grammatik strukturalar va lug‘at tarkibiy qismlarini o‘zida mujassam etgani sababli juda muhim. Bu jarayonda korpuslardan so‘z chastotasi, ibora naqshlari, sintaktik va semantik ma’lumotlar kabi elementlarni to‘plash mumkin.

Talabalar korpuslardan mashina o‘qitish uchun kerakli xususiyatlarni ajratish, tozalash, hamda modelni o‘rgatish uchun tayyorlash jarayonlarini o‘rganadilar. Ular, shuningdek, so‘z vektorlari, n-gramm usullari, va mashina o‘qitish uchun matnni xususiyatlarga aylantirish texnikalari haqida bilimga ega bo‘ladilar. Bu ko‘nikmalar tabiiy tilni qayta ishlashda samarali modellar yaratishga yordam beradi.

12- mavzu. O‘zbek tilida korpus yaratish tajribalari va ularni tahlil qilish hamda boshqarish

O‘zbek tilida korpus yaratish - bu tilshunoslik va tabiiy tilni qayta ishlash sohasida yangi imkoniyatlarni ochib beruvchi murakkab jarayon. O‘zbek tilidagi korpuslar cheklangan bo‘lgani uchun, ularni yaratish va rivojlantirishga katta ehtiyoj bor. Bu jarayonda matnlarni yig‘ish, tahrirlash va strukturalash asosiy bosqichlar hisoblanadi. Hozirgacha O‘zbek tilida korpus yaratishga qaratilgan bir qancha ilmiy loyihalar amalga oshirilgan, ular O‘zbek tilining lug‘at tarkibi, grammatikasi va sintaktik tuzilishini o‘rganishga xizmat qiladi.

Studentlar ushbu mavzuda O‘zbek tilida korpus yaratishning asosiy qiyinchiliklari va usullarini o‘rganadilar. Jumladan, matn to‘plash jarayonlari, yozuv tizimining o‘ziga xosliklari, dialektlar va uslubiy xususiyatlar kabi masalalar haqida bilimga ega bo‘ladilar. Shuningdek, korpusni tozalash, belgilarni normallashtirish, morfologik va sintaktik belgilash jarayonlari bilan tanishadilar. Bu bilimlar O‘zbek tilida mustahkam va funksional korpus yaratish imkonini beradi.

13- mavzu. Korpuslardan mashina tarjimasi uchun foydalanish

Korpuslardan mashina tarjimasi uchun foydalanish - bu tabiiy tilni qayta ishlash (NLP) sohasidagi muhim amaliyot bo‘lib, u ikki yoki undan ortiq til o‘rtasidagi tarjimalarni avtomatlashtirish uchun zarur bo‘lgan ma’lumotlarni taqdim etadi. Mashina tarjimasi uchun parallel korpuslar (bir xil mazmundagi matnlarning ikki yoki undan ortiq tildagi nusxalari) ayniqsa foydalidir, chunki ular so‘z va iboralarning bir tilidan ikkinchisiga qanday tarjima qilinishini o‘rganish imkonini beradi. Bu usul, xususan, neyron tarmoqlarga asoslangan mashina tarjima tizimlarida keng qo‘llanadi.

Studentlar ushbu mavzuda parallel korpuslar bilan ishlash, ularni to‘plash va tahrirlash, tarjima sifatini oshirish uchun modelni o‘rgatish jarayonlarini o‘rganadilar. Shuningdek, segmentatsiya, so‘z birikmalarining ekvivalentlarini topish, grammatik moslik va kontekstual ma’nolarni saqlash kabi muhim texnikalar haqida bilimga ega bo‘ladilar. Bu ko‘nikmalar mashina tarjima tizimlarini yanada aniqroq va tabiiyroq qilishga xizmat qiladi.

14- mavzu. Til korpuslarida sozlamalar va izlash imkoniyatlari

Til korpuslarida sozlamalar va izlash imkoniyatlari - bu foydalanuvchilarga katta hajmdagi matnlar orasida kerakli ma’lumotlarni tez va samarali topishga yordam beruvchi funksiyalardir. Korpuslarda murakkab qidiruv mexanizmlari mavjud bo‘lib, ular orqali

Studentlar ushbu mavzuda korpuslar qidiruv tizimlarini, ular yordamida so‘z chastotasi, kontekstual qo‘llanilishi va grammatika naqshlarini qanday topish mumkinligini o‘rganadilar. Shuningdek, ma’lum sozlamalar orqali ma’lumotlarni filtr qilish, ya’ni maxsus kategoriyalar (masalan, mavzuga, uslubga yoki lahjaga qarab) bo‘yicha qidirish usullari bilan tanishadilar. Bu imkoniyatlar til tahlili, lingvistik tadqiqotlar va NLP dasturlarini yaratishda katta yordam beradi.

15- mavzu. So‘z birikmalarini tahlil qilish usullari va til korpuslarida variantli so‘z shakllari

So‘z birikmalarini tahlil qilish usullari - bu matnlarda tez-tez birga ishlatiladigan so‘zlar kombinatsiyalarini o‘rganish uchun qo‘llaniladigan usullardir. So‘z birikmalarining tahlili tilning leksik va sintaktik tuzilishini tushunishga yordam beradi, shuningdek, u tabiiy tilni qayta ishlash (NLP), mashina tarjimasi, va ma’nolarni aniqlash tizimlarida muhim o‘rin tutadi. Bu usullar yordamida so‘zlar orasidagi bog‘lanishlar, kontekstual ma’nolar va ibora naqshlari o‘rganiladi.

Studentlar ushbu mavzuda n-gramm (ikki so‘zli, uch so‘zli va undan ko‘p so‘zli kombinatsiyalar), chastotaviy tahlil, kollokatsiya, va o‘zaro axborot (mutual information) kabi asosiy usullarni o‘rganadilar. Shuningdek, ular korpuslardan foydalangan holda so‘z birikmalarini tahlil qilish, kombinatsiyalarning semantik xususiyatlarini aniqlash va kontekstual qo‘llanishiga qarab tahlil qilishni o‘zlashtiradilar. Bu ko‘nikmalar tilning tabiiy tuzilishini tushunishga va yanada samarali lingvistik tizimlar yaratishga imkon beradi.

Til korpuslarida variantli so‘z shakllarini aniqlash va tahlil qilish, tilning boyligini, ijtimoiy va kontekstual farqlarni tushunishga yordam beradi. Bunday shakllar til o‘rganishda, so‘z birikmalarini tahlil qilishda va grammatika tuzilmalarini o‘rganishda muhim ahamiyatga ega.

Talabalar bu mavzuda variantli so‘z shakllarining tahlilini, ularning kontekstdagi o‘zgarishini va qaysi faktorlar ularni shakllanishiga ta’sir qilishini o‘rganadilar. Ushbu tahlil, shuningdek, morfologik qayta ishlash va so‘zlarni normalizatsiya qilishda yordam beradi. Korpuslar yordamida variantli shakllarni tahlil qilish, talabalarni tilning nozik nuanslarini, sinonimlar va antonimlar o‘rtasidagi farqlarni va ularning ko‘plab kontekstlarda qanday ishlatilishini aniqlashga yo‘naltiradi. Bu jarayonlar, tilshunoslik, NLP va boshqa lingvistik tadqiqotlar uchun muhimdir.

16- mavzu. Matnni qayta ishlashda til korpuslari Til korpuslarining huquqiy va etik muammolari

Matnni qayta ishlashda til korpuslari - bu tabiiy tilni qayta ishlash (NLP) jarayonlarida foydalanuvchilarga haqiqiy matnlar asosida ko‘plab til qoidalari va strukturalarini o‘rganishga imkon beruvchi resurslardir. Korpuslar orqali turli xil tillarda yozilgan matnlar to‘plamlarini taqdim etish, ularni tahlil qilish va modellar yaratish mumkin. Bu jarayonda korpuslar, avtomatik tarjima, matn tasnifi, sentiment tahlili, so‘z birikmalarini aniqlash va boshqa ko‘plab ilovalarda asosiy o‘rin tutadi.

Studentlar ushbu mavzuda matnni qayta ishlash jarayonida korpuslardan qanday foydalanishni, ularni qanday to‘plash, belgilash va tahlil qilishni o‘rganadilar. Shuningdek, ular korpuslardan foydalanish orqali morfologik, sintaktik va semantik tahlil, jumlalarni segmentatsiya qilish va til modellari yaratish kabi vazifalarni qanday bajarish mumkinligini o‘zlashtiradilar. Bu ko‘nikmalar NLP sohasidagi amaliyot va tadqiqotlar uchun muhimdir, chunki ular yuqori sifatli til tizimlarini yaratishga yordam beradi.

Til korpuslarini yaratish va ulardan foydalanish jarayonida bir qator huquqiy va etik muammolar paydo bo‘lishi mumkin. Ushbu muammolarni aniqlash va hal etish tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohasida muhimdir.

1. **Mualliflik huquqlari**: Korpuslarda ishlatiladigan matnlar ko‘pincha mualliflik huquqiga ega bo‘lgan materiallardir. Korpuslarni yaratishda yoki ulardan foydalanishda mualliflik huquqlarini buzmaslik uchun tegishli ruxsatnomalar olish zarur.
2. **Maxfiylik va shaxsiy ma'lumotlar**: Korpuslar, ayniqsa, ijtimoiy tarmoqlardan olingan ma'lumotlar, shaxsiy ma'lumotlarni o‘z ichiga olishi mumkin. Shaxsiy ma’lumotlar va maxfiylikni himoya qilish muhim ahamiyatga ega bo‘lib, foydalanuvchilar ma’lumotlarning qayerdan olinganini va qanday ishlatilishini bilishlari kerak.
3. **Nohaqlik va kamsitish**: Korpuslar turli ijtimoiy guruhlar, gender, yoki etnik identifikatsiyalarga nisbatan nohaq yoki kamsituvchi til namoyon etishi mumkin. Korpuslarni tahlil qilishda va modellarni yaratishda bunday muammolarga e’tibor berish zarur.
4. **Ommaviy foydalanish va qiyinchiliklar**: Korpuslardan foydalanishning qanday shartlari borligi haqida aniq va ochiq ma’lumot berish, ularning qayerda va qanday maqsadda foydalanilishini belgilash muhimdir.

Studentlar ushbu huquqiy va etik muammolarni o‘rganish orqali til korpuslarini yaratish va ulardan foydalanishda zarur bo‘lgan mas'uliyatli yondashuvni tushunadilar. Bunday bilimlar ularni etika va qonun doirasida ishlashga tayyorlaydi, bu esa lingvistik tadqiqotlar va NLP ilovalari uchun muhimdir.

17- mavzu. Mashhur til korpuslari va ularning xususiyatlari Brown Corpus

Til korpuslari, tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohalarida keng qo‘llaniladigan resurslardir. Ularning bir nechtasi mashhur bo‘lib, har biri o‘ziga xos xususiyatlarga ega. Quyida ba'zi mashhur til korpuslari va ularning xususiyatlari keltirilgan: Brown Corpus, Penn Treebank, Corpus of Contemporary American English (COCA), ruscorpora.

Ushbu mashhur til korpuslari, talabalar va tadqiqotchilar uchun turli xil tilshunoslik va NLP vazifalarini bajarishda muhim resurs bo‘lib xizmat qiladi. Har bir korpusning o‘ziga xos xususiyatlari, ulardan qanday foydalanish va tilni tahlil qilish jarayonida qanday qo‘llanilishi haqida bilish, tilshunoslik va informatika sohasidagi ko‘nikmalarni rivojlantirishga yordam beradi. Brown Corpus - bu tabiiy tilni qayta ishlash va tilshunoslik tadqiqotlarida keng qo‘llaniladigan eng mashhur korpuslardan biridir. 1960-yillarda (an'anaga ko'ra 1961 yilda) R. Hudzon va uning hamkasblari tomonidan yaratilgan. U ingliz tilidagi 1 million so‘zni o‘z ichiga oladi va turli janrlarda yozilgan matnlarni taqdim etadi. Brown Corpus, ingliz tilidagi eng muhim til resurslaridan biri bo‘lib, tilshunoslik, leksikologiya, stilistika va tabiiy tilni qayta ishlash sohalarida qimmatli manba hisoblanadi. U talabalarga va tadqiqotchilarga tilni chuqurroq o‘rganishga va yangi tadqiqotlar o'tkazishga yordam beradi.

18- mavzu. Til korpuslarini tahlil qilish uchun dasturiy vositalar Til korpuslarni yaratish

Til korpuslarini tahlil qilish jarayonida foydalaniladigan dasturiy vositalar tilshunoslar va tadqiqotchilarga matnlarni yanada samarali va aniq tahlil qilish imkonini beradi. Quyida

	mashhur dasturiy vositalar va ularning asosiy xususiyatlari keltirilgan: NLTK (Natural Language Toolkit), spaCy, Gensim, Sketch Engine, AntConc, Treetagger, extRazor. Ushbu dasturiy vositalar, talabalar va tadqiqotchilarga til korpuslarini tahlil qilishda yordam beradi, bu esa ularga tilning tuzilishi, leksik xususiyatlari va grammatik qoidalarini yanada chuqurroq tushunishga yordam beradi. Talabalarni til korpuslarini yaratish jarayonlari, ularning ahamiyati va foydalanish usullari bilan tanishtirish, shuningdek, tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohalarida korpuslardan qanday foydalanishni o'rganishga yo'naltirish. Ushbu mavzu orqali talabalar til korpuslarini yaratish jarayonini to'liq tushunishga va kelajakda o'z tadqiqotlarida ushbu resurslardan samarali foydalanishga tayyorlanadilar. **III. Mustaqil ta'lim*** **III.1. Mustaqil ta'lim va mustaqil ishlar** Talabalarni mustaqil ta'lim shaklini tashkil etishga qo'yilgan talablar O'zbekiston Respublikasi oliy ta'lim, fan va innovatsiyalar vazirligining 2024-yil 29-apreldagi 136-sonli "Oliy ta'lim muassasalari talabalari mustaqil ta'limini tashkil etish bo'yicha namunaviy tartibni tasdiqlash to'g'risida"gi buyrug'i asosida ishlab chiqilgan. Mustaqil ta'limni baholash semestr davomida berilgan topshiriq asosida bajarilgan ishlarni HEMISda ilova qilish, shuningdek, oraliq va yakuniy test va savollarga javob berish asosida oshiriladi. **"Til korpuslari "** fanidan mustaqil ta'lim uchun tavsiya etiladigan mavzular: Til korpuslariga kirish Til korpuslarining turlari Korpus lingvistikasining asosiy tushunchalari Korpuslar yaratishning bosqichlari Ma'lumot yig'ish usullari Korpus tuzilishi va formatlari Korpusni tozalash va normalizatsiya qilish Annotatsiya va markup Korpusda morfologik belgilash Sintaktik annotatsiya Semantik belgilash Paralel korpuslar Til korpusida tez-tez uchraydigan iboralar Morfologik va sintaktik analizatorlar bilan ishlash Korpuslardan kontekst olish va foydalanish Keyingi so'zlarni prognoz qilish Til o'rganishda korpuslardan foydalanish Talabalarni mustaqil ta'lim shaklini tashkil etishga qo'yilgan talablar O'zbekiston Respublikasi oliy ta'lim, fan va innovatsiyalar vazirligining 2024-yil 29-apreldagi 136-sonli "Oliy ta'lim muassasalari talabalari mustaqil ta'limini tashkil etish bo'yicha namunaviy tartibni tasdiqlash to'g'risida"gi buyrug'i asosida ishlab chiqilgan. Mustaqil ta'limni baholash semestr davomida berilgan topshiriq asosida bajarilgan ishlarni HEMISda ilova qilish, shuningdek, oraliq va yakuniy test va savollarga javob berish asosida oshiriladi. * *Izoh: Mustaqil ta'lim ishi mavzulari fan doirasida o'zgarishi mumkin.*
4.	**V. Fan o'qitilishining natijalari (shakllanadigan kompetensiyalar)** 1. Til korpuslari **bilimlari:**

	○ Talabalar til korpuslari haqida nazariy bilimlarni o‘zlashtiradilar, ularning tuzilishi, maqsadi va funksiyalarini tushunadilar. ○ Til korpuslari orqali tilning strukturasi, leksikasi va sintaksisi haqida chuqurroq bilim oladilar. 2. **Ma'lumotlarni tahlil qilish qobiliyati:** ○ Talabalar til korpuslaridan ma'lumotlarni yig‘ish va tahlil qilish ko‘nikmalarini rivojlantiradilar. ○ Tahlil natijalarini interprete qilish va ulardan xulosa chiqarish qobiliyatini shakllantiradilar. 3. **Axborot texnologiyalaridan foydalanish:** ○ Talabalar turli dasturiy ta'minotlar va vositalardan foydalanishni o‘rganadilar, masalan, til korpuslarini yaratish, tahlil qilish va natijalarni vizualizatsiya qilishda. ○ Korpuslarni yaratish va boshqarish uchun zamonaviy axborot texnologiyalarini qo‘llash ko‘nikmalarini egallaydilar. 4. **Izlanish va ilmiy tadqiqot ko‘nikmalari:** ○ Talabalar til korpuslari asosida ilmiy tadqiqotlarni amalga oshirish ko‘nikmalarini rivojlantiradilar. ○ Til korpuslaridan foydalanib, o‘z mustaqil izlanishlarini olib borish imkoniyatiga ega bo‘ladilar. 5. **Tanqidiy fikrlash:** ○ Talabalar korpusdan olingan ma'lumotlarni tahlil qilish orqali tanqidiy fikrlash qobiliyatini oshiradilar. ○ Ular tahlil jarayonida olingan natijalarni baholash va muhokama qilish ko‘nikmalarini rivojlantiradilar.
5.	**VI. TA’LIM TEXNOLOGIYALARI VA METODLARI**: **Interaktiv ta’lim metodlari:** • **Multimedia materiallar:** Talabalar til korpuslari haqida ko‘rgazmali ma'lumotlarni olish uchun video, audio va grafik materiallardan foydalanishadi. • **Interaktiv platformalar:** Talabalar turli onlayn platformalarda birgalikda ishlashlari, masalan, Google Docs, kahoot.com kabi vositalar orqali til korpuslarini tahlil qilishlari mumkin. **O‘qitish metodlari:** • **Tadqiqot asosidagi ta'lim:** Talabalar til korpuslari ustida mustaqil tadqiqotlar olib borish orqali amaliy tajribaga ega bo‘ladilar. • **Loyihaga asoslangan o‘qitish:** Talabalar guruhlar bo‘lib, til korpuslari yaratish yoki tahlil qilish loyihalarini amalga oshiradilar. Bu ularga jamoada ishlash va muammolarni birgalikda hal qilish ko‘nikmalarini rivojlantirishga yordam beradi.
6.	**VII. Kreditlarni olish uchun talablar:** Topshiriqni muvaffaqiyatli bajarish uchun asosiy ko‘nikmalar:

	Nazariy bilimlar: zarur ma'lumotlarni to‘plash, ularni tahlil qilish va anglash qobiliyatiga ega bo‘lishlari kerak. Yangi ma'lumotlarni o‘rganish va ulardan foydalanish qobiliyati, shu jumladan, til korpuslari bo‘yicha ilmiy tadqiqotlar olib borish. **Amaliy ko‘nikmalar:** Zamonaviy texnologiyalardan (kompyuter dasturlari, ilovalar) foydalanish qobiliyati, masalan, til korpuslarini yaratish va tahlil qilishda dasturiy ta'minotlardan foydalanish. Laboratoriya yoki amaliy mashg‘ulotlarda tajriba orttirish, talabalar o‘z ko‘nikmalarini amaliyotda qo‘llashlari zarur.
7.	**VIII. ASOSIY VA QO‘SHIMCHA O‘QUV ADABIYOTLAR HAMDA AXBOROT MANBALARI** **Asosiy adabiyotlar** 1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика:учебник 3-изд., перераб.- СПб. Изд-во. С.-Петерб. ун-та, 2020.-234 с. ISBN 978-5-288-05997-1 2. Jonathan Dunn NATURAL LANGUAGE PROCESSING FOR CORPUS LINGUISTICS University Printing House, Cambridge CB2 8BS, United Kingdom One Liberty Plaza, 20th Floor, New York, NY 10006, This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press. First published 2022 A catalogue record for this publication is available from the British Library. ISBN 978-1-009-07443-8 Paperback ISSN 2632-8097 (online) ISSN 2632-8089 (print) 3. Joanna Baumgart Corpus Linguistics and Cross-Disciplinary Action Research A Study of Talk in the Mathematics Classroom First published 2022 by Routledge 2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN and by Routledge 605 Third Avenue, New York, NY 10158 Routledge is an imprint of the Taylor & Francis Group, an informa business **Qo'shimcha adabiyotlar** 4. Mirziyoyev Sh.M. Hozirgi zamon va Yangi O‘zbekiston. - Toshkent: O‘zbekiston, 2024. 5. Mirziyoyev Sh.M. Yangi O‘zbekistonda taraqqiyot strategiyasi asosida demokratik islohotlar yo‘lini qat’iy davom ettiramiz. 6-jild. - Toshkent: O‘zbekiston, 2023 6. Апресян Ю.Д. Идеи и методы современной структурной лингвистики. - М., 2006. 7. Лапшин В.А. Лекции по математической лингвистике. М.: Научный мир, 2010. 248 с.

	8. Арапов М. В., Херц М. М. Математические методы в исторической лингвистике. М., 2009. 9. Грудева Е.В. Корпусная лингвистика [Электронный ресурт]: учеб. Пособие/ 2- изд., стер.- М.: Флинта, 2012.-165 с. 10. Захаров В.П., Богданова С.Ю. З-38 Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с. ISBN 978-5-88267-316-0 **Scopusdan maqolalar** 11. The Impact of Word Splitting on the Semantic Content of Contextualized Word Representations Aina Garí Soler, Matthieu Labeau, Chloé Clavel Transactions of the Association for Computational Linguistics (2024) 12: 299–320. https://doi.org/10.1162/tacl_a_00647 https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00647/120475/The-Impact-of-Word-Splitting-on-the-Semantic 12. Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization Zhixue Zhao, George Chrysostomou, Miles Williams, Nikolaos Aletras Transactions of the Association for Computational Linguistics (2024) 12: 1163–1181. https://doi.org/10.1162/tacl_a_00695 https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00695/124459/Investigating-Hallucinations-in-Pruned-Large **Axborot manbalar:** **https://ruscorpora.ru/** **https://collins.co.uk/** **https://www.ucl.ac.uk/english-usage/ice.htm** **http://helmer.aksis.uib.no/icame/brown/bcm.html**
8.	Fanning o'quv dasturi Tarjimashunoslik, tilshunoslik va xalqaro jurnalistika oliy maktabida ishlab chiqilgan va 2025 yil "18" 06 dagi 24 -sonli bayonnomasi bilan ma'qullangan. O'quv dasturi Toshkent davlat sharqshunoslik universiteti Kengashining 2025 yil "28" 06 17 sonli bayoni bilan tasdiqlangan.
9.	**Fan/modul uchun ma'sullar:** **1.** Allanyazov R.B.– TDShU Tarjimashunoslik, tilshunoslik va xalqaro jurnalistika oliy maktabi o'qituvchisi
10.	**Taqrizchilar:** 1. Usmanova Shoira Rustamovna - Tarjimashunoslik, tilshunoslik va xalqaro jurnalistika oliy maktabi professori, f.f.d. 2. N.Abdurahmonova - O'zMU, Kompyuter lingvistikasi va amaliy tilshunoslik"kafedrasi professori, f.f.d