

KORPUS LINGVISTIKASI

O'ZBEK TILI MILLIY KORPUSINING DASTURIY TA'MINOT STRUKTURASI VA VAZIFALARI

Qarshiyev Abduvali Berkinovich

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Samarqand filiali professori,
Fizika-matematika fanlari nomzodi,

Karimov Suyun Amirovich

Samarqand davlat universiteti professori,
Filologiya fanlari doktori,

Tursunov Muhammadsolih Sa'din o'g'li

muhammadsolih927@gmail.com

Samarqand davlat universiteti mustaqil izlanuvchisi

Annotatsiya: Lingvistik korpus yordamida tilni tahlil qilish va o'rganish orqali uning qonuniyatlarini tadqiq etish tilshunos sohasida ulkan yutuq deb aytish mumkin. Ushbu maqolada o'zbek tili milliy korpusining dasturiy ta'minoti strukturasi va vazifalari keltirilgan. Mavjud korpuslar tahlilidan ko'rish mumkinki, milliy korpus tilning qonuniyat va xususiyatlarini iloji boricha ko'p taqdim eta olishi va tarkibi bo'yicha muvozanatlashgan bo'lishi kerak. Matnlarning bir turi tildagi jami matnlar ichida qanday hajmiy ulushga ega bo'lsa, ana shunday miqdor milliy korpusda ham saqlanib qolishi kerak.

Kalit so'zlar: *o'zbek tili milliy korpusi, dasturiy ta'minot, model, tokenayzer, matnlar bazasi, lug'at, grammatik ma'lumot.*

O'zbek tili milliy korpusi modeli va uni yaratish uchun bajariladigan ishlar bosqichlaridan bog'liq holda dasturiy ta'minot strukturasi ishlab chiqildi. Dasturiy ta'minot ikki qismdan iborat bo'ladi (1-rasm):

1. korpusni yaratishga mo'ljallangan dasturlar;
2. korpusdan foydalanishga xizmat qiladigan dasturlar.



1-rasm. Dasturiy ta'minot tarkibi

Korpusni yaratishga mo'ljallangan dasturlar

Korpusni yaratishga mo'ljallangan dasturlar matnlar bazasini shakllantirish, korpus lug'atini shakllantirish va tahrirlash, matnlarni razmetkalash kabi vazifalarni bajarashi kerak (2-rasm).



2-rasm. Korpus yaratish dasturlari tarkibiy qismlari

Matnlar bazasini shakllantirish dasturlari

Matnlar bazasini shakllantirishda quyidagi ishlar bajarilishi zarur:

- matnlarni raqamlashtirish, ularni tahrirlash;
- matn bo'yicha ma'lumotlarni faylga yozish;
- matnni bazaga kiritish.

Matnlarni raqamlashtirish vositalarini ularning dastlabki holati qaysi manbada (qog'oz, *.pdf formatdagi fayl) ekanligidan bog'liq holda tanlaymiz.

Qog'oz manbadagi matnni raqamlashtirish uchun dastlab uni skanerlab olamiz. Bunda skanerlash apparati va skanerlash dasturlari ishlatiladi. Hosil bo'lgan elektron matn *MS Word* dasturi yordamida *.docx formatga o'giriladi, so'ngra asosiy manba bilan solishtirib, sinchiklab tekshiriladi, xatolar to'g'rilanadi va saqlanadi.

Matn *.pdf formatdagi faylda berilgan bo'lsa, u dasturiy vositalar yordamida *.docx formatga o'tkaziladi. So'ngra asosiy manba bilan solishtirib, sinchiklab tekshiriladi, xatolar to'g'rilanadi va saqlanadi.

Matnlar bazasini yaratishga doir ishning keyingi bosqichlarini *Dast_MtnBaza* nomli dastur amalga oshiradi. Bu dastur saqlangan fayllar asosida korpusning matnlar bazasini shakllantiradi. Matnlar bazasi kompyuter xotirasidagi bitta *Matn_Baza* nomli papka va bitta *MeteRazm* nomli metarazmetka ma'lumotlari faylidan iborat bo'ladi. *Matn_Baza* papkada korpus tarkibiga kiritishga qaror qilingan matnlarni o'z ichiga oluvchi *.docs fayllar saqlanadi. *MeteRazm* faylda *Matn_Baza* papkaga kiritilgan har bir faylda mavjud matn haqida metarazmetka ma'lumotlari saqlanadi. Metarazmetka ma'lumotlari matn bo'yicha umumiy ma'lumotlardan iborat bo'lib, quyidagilarni o'z ichiga oladi:

- matn nomi;
- matn muallifiga oid ma'lumotlar: ismu-sharifi, jinsi, tug'ilgan sana va yili va h.k.;
- matn yozilgan vaqt;
- tematika va matn turi;
- janr;
- matn hajmi (so'zlarda).

Dastur ushbu ma'lumotlarni foydalanuvchi tomonidan kiritilishini ta'minlaydi va ularni *MeteRazm* faylga yozib qo'yadi. So'ngra kiritilgan metarazmetka ma'lumotlar asosida maxsus yagona ism shakllantiriladi va matn saqlanib turgan *.docs fayl shu ism bilan *Matn_Baza* papkaga ko'chirib yoziladi.

Korpus lug'atini shakllantirish va tahrirlash

Matnni grammatik (morfologik) razmetklash uchun tilning grammatik lug'ati asos vazifasini bajarishi kerak. Masalan, rus tili milliy korpusi uchun A.A.Zaliznyakning rus tili grammatik lug'ati [1] asos vazifasini bajaradi. Bu lug'at elektron shaklga o'tkazilib, ruscha so'zlarni razmetkalashda ishlatiladi. Lekin o'zbek tili uchun bunday lug'at mavjud emas. Shuning uchun o'zbekcha so'zlarning grammatik elektron lug'atini shakllantirish zaruriyati yuzaga keladi.

Grammatik lug'at yaratish uchun *Gram_Lugat* nomli dastur tuzilgan. Lug'at korpusning matnlar bazasiga kiritilgan matnlar asosida shakllanadi. *Gram_Lugat* dastur korpusning *Matn_Baza* matnlar bazasidagi har bir faylni navbat bilan ishlovdan o'tkazadi. Har bir fayl uchun *Gram_Lugat* dastur quyidagi ishlarni amalga oshiradi:

- matnni tokenlash;
- matndagi so'zlarning turli ro'yxatlarini (alfavit-chastotali tartibdagi, chastota-alfavit tartibdagi va ters ro'yxat) tuzish;
- so'zlarni grammatik lug'atga kiritish;
- grammatik lug'atni varaqlab ko'rish va tahrirlash.

Matnni tokenlash

Matnni avtomatik qayta ishlashda, birinchi navbatda, undagi so'zlarni ajratib olish, yoki boshqacha aytganda, matnni birliklarga bo'laklash masalasi yuzaga keladi. Buning uchun ajratuvchi belgilarni (probel, tinish belgilari va h.k.) o'z ichiga olmagan hamma qisman satrlar matndan ajratilib olinishi lozim. Bu esa tokenlar to'plami bo'ladi [2]. Matnlarni avtomatik qayta ishlashning fundamental

algoritmalaridan biri, berilgan matnni tokenlarga bo‘lib tashlashdan iborat. Algoritm kirishiga matn berilib, chiqishida matndagi tokenlar ro‘yxati olinadi. Bu algoritmni amalga oshiruvchi dasturni tokenayzer deb atashadi. Odatda, tokenlar so‘z shakllari bilan bir xil ma‘noni beradi. Lekin, leksik birliklarni ifodalash uchun “so‘z” emas, balki “token” termini ishlatiladi. Bunga sabab, ba’zi hollarda token sifatida so‘zdan kichikroq birliklar (alohida morfema) yoki so‘zdan kattaroq birliklar (so‘z birikmalari) ishlatilishi mumkin.

Tokenayzer matnni, dastlab, so‘zlar orasidagi probellar (bo‘shliq belgilari) asosida bo‘laklaydi, so‘ngra so‘zlardan tinish belgilari olib tashlanadi. Qisqartmalar (masalan, TATU, BMT, MDH, h.k.) va sana yozuvi (masalan, 09.04.2018) ham token sifatida olinadi [3].

Tokenayzer ishining natijalari quyidagicha bo‘ladi:

1-jadval

№	Berilgan matn	Tokenlar ro‘yhati
1	<i>O‘zbekiston Respublikasi 02.03.1992 kuni BMT ga a‘zo bo‘lgan</i>	<i>O‘zbekiston Respublikasi 02.03.1992 kuni BMT a‘zo bo‘lgan</i>
2	<i>TATU Samarqand filiali 2005 yilda o‘z faoliyatini boshladi</i>	<i>TATU Samarqand filiali 2005 yilda o‘z faoliyatini boshladi</i>

Leksik dekompozisiya matnning avtomatik tahlili uchun fundamental ahamiyatga ega, chunki bu bir qator boshqa algoritmlar uchun asos vazifasini bajaradi.

Matndagi soʻzlarning turli roʻyxatlarini tuzish

Bu yerda maqsad – matnning lugʻatga kiritilishi lozim boʻlgan birliklarini (soʻzlar, soʻz shakllari) ajratib olish, ularning roʻyxatini qilish va roʻyxatni foydalanuvchi talabi bilan turli shakllarda (alfavit-chastotali tartibdagi, chastota-alfavit tartibdagi va ters roʻyxat) taqdim etish. *Gram_Lugat* dasturning bu vazifani bajarishga mos moduli *Suz_Rhati* deb nomlanib, uning kirishida matndan ajratilgan tokenlar roʻyxati boʻlsa, chiqishida soʻz va soʻz shakllarining tartiblangan roʻyxati boʻladi. Bu roʻyxatni shakllantirish uchun dastur gramatik lugʻat fayli *LUGAT*dan foydalanadi. Kirishda berilgan har bir token mavjud mavjud *LUGAT* fayldan izlab koʻriladi. Agar token lugʻatdan topilsa, demak, uni lugʻatga kiritish zaruriyati yoʻq, tashlab yuborish mumkin. Aks holda, yaʼni token lugʻatdan topilmasa, u soʻz yoki soʻz shakli ekanligi tekshiriladi. Soʻz yoki soʻz shakli boʻlgan tokenlar alohida roʻyxatga kiritiladi. Bu roʻyxat foydalanuvchi talabiga mos holda, yo alfavrit-chastotali tartibda, yo chastota-alfavit tartibda, yo ters roʻyxat shaklida saralanadi va ekranga chiqariladi.

Soʻzlarni grammatik lugʻatga kiritish

Bu vazifani bajaruvchi dasturiy modul *Lug_Kirit* deb nomlangan. Lugʻatga yangi birlik kiritish foydalanuvchi tomonidan interaktiv rejimda amalga oshiriladi. Bu jarayonni *Lug_Kirit* dasturi taʼminlaydi. Kompyuter ekranida *Suz_Rhati* dasturi tomonidan chiqarilgan roʻyxat mavjud va unda bitta soʻz boshqa rang bilan ajratib koʻrsatilgan. *Lug_Kirit* dasturi yordamida foydalanuvchi roʻyxat boʻyicha harakatlanib, joriy soʻzni tanlaydi va “Enter” tugmasini bosadi. Natijada ekranda kiritish oynasi paydo boʻladi. Foydalanuvchi bu oynada soʻralgan maʼlumotlarni klaviaturadan terib kiritadi va “Bazaga kiritish” tugmasini bosadi. Shunda tanlangan soʻz va unga mos kiritilgan grammatik xarakteristikalar jamlanib, elektron lugʻat fayli *LUGAT*ga va boshqa yordamchi fayllarga yozib qoʻyiladi. Shundan soʻng foydalanuvchi roʻyxatdan navbatdagi soʻzni tanlaydi va bazaga kiritadi. Shu tariqa,



berilgan ro'yxatdagi barcha so'zlar LUGAT faylga kiritilib, gramatik lug'at qadamma-qadab boyitib boriladi.

Foydalanilgan adabiyotlar:

1. А.А.Зализняк, грамматический словарь русского языка, Москва, 1980.
2. Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) Прикладная и компьютерная лингвистика, Ленанд, 2016. - 320 с. - ISBN 978-5-9710-3472-8.
3. Tursunov M.S., Qarshiyev A.B., Karimov S.A., Development of a modern corpus of computational linguistics, Scopus DOI: 10.1109/ICISCT50599.2020.9351376