

## Matnga grammatik ishlov berish tizimlari

### SO‘Z TURKUMLARINI TEGGLASH USULLARI: MUAMMO VA YECHIMLAR

**Axmedova Xolisa Ilxomovna**

Alisher Navoiy nomidagi Toshkent davlat  
o'zbek tili va adabiyoti universiteti tayanch doktoranti  
[xolisa9029@mail.ru](mailto:xolisa9029@mail.ru)

**Yusupova Dilduza Muhammadqodirovna**

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

**Annotatsiya.** Tabiiy tillarni qayta ishlash jarayonidagi qator muammolar o'z yechimini kutmoqda. Shunday muammolardan biri so'zlarni semantic tahlil qilish muammosidir. Bu muammoni yechish jarayonida turli yondashuvlar, modellarga duch kelamiz.

**Kalit so'zlar:** *Tabiiy tillarni qayta ishlash, so'z turkumlari, Yashirin Morkov model, O'tish ehtimollik, Emission ehtimollik*

**Annotation.** There are a number of problems in the process of processing natural languages. One such problem is the problem of semantic analysis of words. In the process of solving this problem, we encounter different approaches, models.

**Keywords.** *Natural Languages Processing, part of speech, Hidden Morkov model, Transition probabilities, Emission probabilities.*

Tabiiy tillarni qayta ishlash jarayoni (NLP-Natural Languages Processing) ning muammolarini ketma –ket modellashtirish zamon talabidir. Ayniqsa so'z turkumlarini teglash qadimiy va eng mashhur muammolardan hisoblanadi. So'z turkumlarini teglashtirishdagi asosiy maqsadimiz gap tuzilishini modellashtirishdan iborat. Masalan, gapni o'qish va qanday so'z lar ot, olmosh, fe'l, ravish va hokazo vazifasini bajarishini aniqlash nutq teglarining bir qismi hisoblanadi. *Korpus tilshunosligida so'z turkumlarini teglash, grammatik teglash yoki so'z turkumlariga ajratish deb ham ataladi, bu matn (korpus) dagi so'z ni nutqning ma'lum bir qismiga mos keladigan tarzda belgilash, uning ta'rifi va kontekstidan kelib chiqqan holda, ya'ni o'zidan oldin va keyin kelgan va o'zaro bog'liq ibora, jumla yoki paragrafdagi so'z lar.*

So'z turkumlarini teglash shunchaki so'z larni so'z turkumlariga solishtirishdan ko'ra ancha murakkabroq. Chunki so'z turkumlarini teglash oson narsa emas. Bitta so'z uchun turli xil kontekst asosida turli xil jumalarda turli so'z turkumi vazifasida bo'lishi mumkin. Shuning uchun ST(so'z turkumalari) teglari

uchun umumiy qoidani hosil qilish mumkin emas.

Soʻz lar koʻpincha soʻz turkumi sifatida turli maʼnolarda uchraydi. Masalan *Stol ustida olma turibdi.*

*Oʻglim pichoqni qoʻlingga olma.*

Birinci gapda **olma** soʻz i keyingisiga qaraganda boshqa maʼno anglatadi, lekin eng muhimi u ot soʻz turkumiga mansub, ikkinchi gapda esa feʼl. Soʻz larni soʻz turkumlari teglari bilan ajratilsa, soʻz ni maʼnosini ajratib koʻrsatish mumkin. Soʻz lar omonim soʻz lar boʻlgandagina uni gapda aynan qay maʼnoda kelganini aniqlash mumkin. Buni aniqlashda bu soʻz bilan brikkan boshqa soʻz larning soʻz turkumlari teglanishi muhim ahamiyat kasb etadi. Tilimizdagi barcha soʻz larga ST teglarini qanday qilib berish mumkinligini koʻrib chiqamiz.

### Soʻz turkumlari teglarining turlari

ST teglari 2 ta oʻziga xos guruhga boʻlinadi:

- Qoidalarga asoslangan ST teglari
- Stoxastik ST teglari

ST teglash dastlab ingliz tili korpusi yaratilishida qoʻllanilgan. Dastlabki va eng koʻp ishlatiladigan inglizcha ST-teglardan biri boʻlgan E. Brillning<sup>3</sup> qoidalarga asoslangan algoritmlarida qoʻllanilgan. Eng avval qoidalarga asoslanib teglash nimani anglatishini qisqacha koʻrib chiqamiz.

**Qoidalarga asoslangan teglash.** Soʻz turkumlarini avtomatik teglash bu tabiiy tilni qayta ishlash jarayoni boʻlib, unda statistik metodlar bilan qoidalarga asoslangan usullardan koʻra koʻproq muvaffaqiyatga erishilgan. Odatda qoidalarga asoslangan yondashuvlar nomaʼlum yoki noaniq soʻz larga teglar belgilash uchun kontekstli maʼlumotlardan foydalanadi. Soʻz ning lisoniy xususiyatlarini ajratish, oldingi, keyingi soʻz ini va boshqa jihatlarini tahlil qilish orqali amalga oshiriladi. Masalan, oldingi oʻrindagi soʻz *kitob* boʻlsa, u holda bu soʻz ot boʻlishi kerak. Shu kabi boshqa soʻz turkumlari uchun ham qoidalar keltirish mumkin:

*Agar nomaʼlum X soʻz idan oldin aniqlovchi va undan keyin ot kelsa, u holda bu soʻz sifat deb belgilanadi.*

Yuqoridagi keltirilgan qoidalar toʻplami orqali aniqlash juda murakkab jarayon boʻlib, u umuman **doimiy** emas. Shunday ekan, bizga qandaydir avtomatik usul zarur. Brillning teglari - bu maʼlumotlarni oʻrganib chiqadigan va ST teglarining xatolarini minimallashtiradigan qoidalar toʻplamini aniqlaydigan qoidalarga asoslangan teglari. Bu erda Brillning teglarida eʼtibor qaratish kerak boʻlgan eng muhim jihat shundaki, qoidalar qoʻlda ishlanmaydi, aksincha taqdim etilgan korpus yordamida aniqlanadi.

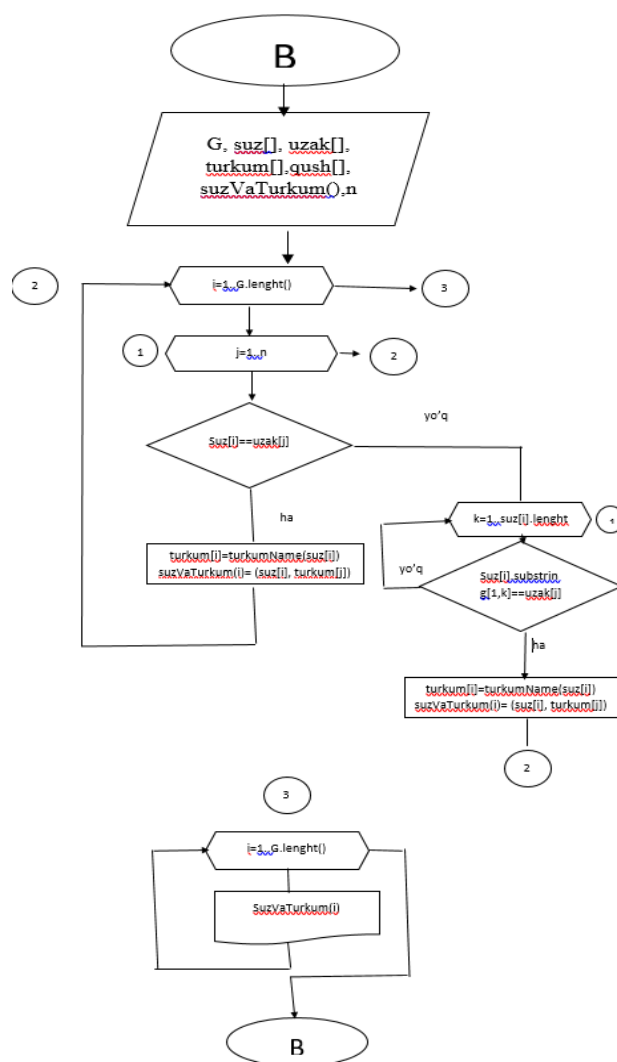
Endi esa Stoxastik ST teglash usulini koʻrib chiqamiz.

**ST teglashning Stoxastik usuli.** "Stochastic tagging" atamasi ST tegi muammosiga turli xil yondashuvlarni nazarda tutishi mumkin. Har qanday holatda chastota yoki ehtimollikni oʻz ichiga olgan har qanday model toʻgʻri stoxastik

<sup>3</sup> E.Brill- ingliz korpusi yaratishfa oʻzining algoritmlari bilan ishtirok etgan olim.

model deb belgilanadi. Eng oddiy stoxastik teglash so'z larni faqatgina so'z ning ma'lum bir teg bilan paydo bo'lish ehtimoli asosida ajratib turadi. Boshqacha qilib aytadigan bo'lsak, ushbu so'z bilan uchrashi mumkin bo'lgan so'z lar to'plamida eng ko'p uchraydigan teg bu so'z ning noma'lum misoliga berilgan tegdir.

So'z larning chastotali yondashuviga alternativ teglar ketma-ketligining yuzaga kelish ehtimolini hisoblashdir. Bu ba'zan **n-gramm** yondashuvi deb ataladi, bunda ma'lum bir so'z uchun eng yaxshi teg avvalgi **n** teglar bilan yuzaga kelish ehtimoli bilan belgilanadi. Ushbu yondashuv ilgari aniqlanganidan ancha mantiqiydir, chunki u alohida so'z lar uchun teglarni kontekst asosida ko'rib chiqadi. Stoxastik teglashda kiritilishi mumkin bo'lgan keyingi murakkablik darajasi avvalgi ikkita yondashuvni ikkala teglar ketma-ketligi ehtimoli va so'z chastotasini o'lchovlari yordamida birlashtiradi. Bu esa Yashirin Markov Modeli deb ataladi. Ingliz tili grammatikasida bu modeldan foydalanish juda qulay va samarali. Chunki ingliz tilida so'zlarga qo'shiladigan lug'aviy va sintaktik shakl yasovchi qo'shimchalar o'rnida predloglar uchraydi, o'zbek tilida esa aksincha. O'zbek tilidagi gaplardagi so'z turkumlariga ajratishni quyidagi blok sxemada ko'rish mumkin.



Bu yerda G- berilgan gap, s[]-G ning so'zlaridan iborat massiv, uzak[]-tilimizda mavjud o'zak so'z lardan iborat massiv, qush[]-so'z larga qo'shiluvchi lug'aviy va sintaktik sahkl yasovchi qo'shimchalardan iborat massiv, suzVaTurkum()-so'z va uning ST tegidan iborat to'plam.  
Masalan.

*Biz hozir ilmiy tadqiqot mavzuyimiz ustida ishlayapmiz.*

Bu gapni yuqoridagi blok sxema yordamida teglashtirsak quyidagi ko'rinishni oladi

*suzVaTurkum([biz,olmosh], [hozir, ravish], [ilmiy, sifat],[tadqiqot, ot],[mavzuyimiz, ot],[ustida, ko'makchi],[ishlayapmiz, fe'l])*

So'z larni so'z turkumlariga ajratib olish jarayonida muammolarga duch kelamiz, ya'ni bitta so'z turli so'z turkumlari vazifasida kelishi mumkin. Bu muammoni yechishda biz Yashirin Morkov modelidan foydalanamiz, buning uchun esa biz so'z lar va ulardan keyin, oldin uchraydigan so'z lar to'plamiga ehtiyoj sezamiz. Yashirin Morkov modeli shu so'z larning dastlabki uchrash holatlari va ularning ehtimolliklari asosida joriy so'z ning qaysi so'z turkumi vazifasida kelishini aniqlab beradi [1]. Bu yerda ikki xildagi ehtimollik kerak bo'ladi, bu ehtimolliklar massiv ko'rinishida hosil bo'ladi va Emmission va O'tish ehtimolliklari deb nomlanadi. Emmission ehtimollik bu har bir so'z ning qaysi so'z turkumi ekanligi ehtimolligi bo'lsa O'tish ehtimolligi ikkita so'z ning ketma-ket kelish ehtimolligi – bu esa so'z larning semantik tahlilida yordam beradi. Yuqorida keltirilgan namunani hisoblashni talab qiladigan quyidagicha ehtimolliklar mavjud:

$P(biz|olmosh), P(hozir/ravish), P(ilmiy/sifat), P(tadqiqot/ot), P(mavzuyimiz/ot), P(ustida| ko'makchi), P(ishlayapmiz| fe'l)$

Bu Emmission ehtimolliklarida, so'zning ya'ni  $P(hozir/ravish)$  hozir so'zining ravish ekanligi ehtimoli qanday bo'lishi mumkin. Shuni esda tutish kerakki, bu HMM yordamida ST teglash muammosini qanday modellashtirish to'g'risida juda oddiy tushuncha berish uchun muammoni rasmiy modellashtirishdir. Agar bizda holatlar to'plami bo'lsa, ketma-ketlik ehtimolini hisoblashimiz mumkin edi. Ammo bizda holatlar yo'q. Bizda faqat kuzatishlar ketma-ketligi bor. Shuning uchun ushbu model Yashirin Markov modeli deb nomlanadi - chunki vaqt o'tishi bilan haqiqiy holatlar yashiringan.



### **Foydalanilgan adabiyotlar:**

1. Michael Collins “Tagging with Hidden Markov Models” 2011.
2. Divya Godayal, An introduction to part-of-speech tagging and the Hidden Markov Model. 2018.
3. Кутузов А.Б. Корпусная лингвистика. – (Электрон ресурс): Лицензия Creative commons Attribution Share-Alike 3.0 Unported (Электрон ресурс)- // lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf;
4. Недошивина Е.В. Программы для работы с корпусами текстов: обзор основных корпусных менеджеров. Учебно-методическое пособие. – Санкт-Петербург. –2006. 26 с.;
5. Mengliyev B. va b. O‘zbek tilining milliy korpusi // Ma’rifat. – 26/04/2018.