

O‘ZBEK TILI KORPUSIDA TEGGLASH METODLARI VA USULLARI

O‘dayeva Shahnoza Shuxratovna

BuxDU mustaqil izlanuvchisi

Annotatsiya. Ushbu maqolada o‘zbek tili korpusida teglash metodlari va usullari xususida so‘z yuritiladi. O‘zbek milliy korpusida teglashdan foydalanishning o‘ziga xos jihatlari mavjud. Qayd etish lozimki, teglash jarayoni ma’lum qoidalarga asoslanib, undan foydalanish orqali tilda mavjud elementlarni modellashtirish mumkin.

Abstract. This article talks about tagging methods and techniques in the Uzbek language corpus. The use of tagging in the Uzbek national corpus has its own characteristics. It should be noted that the tagging process is based on certain rules, and by using it, it is possible to model the elements present in the language.

Аннотация. В данной статье рассказывается о методах и приемах разметки в корпусе узбекского языка. Использование тегов в узбекском национальном корпусе имеет свои особенности. Следует отметить, что процесс разметки основан на определенных правилах, и с их помощью можно моделировать элементы, присутствующие в языке.

Kalit so‘zlar: *korpus lingvistikasi, teglash, modellashtirish, qolip, asos, til birliklari.*

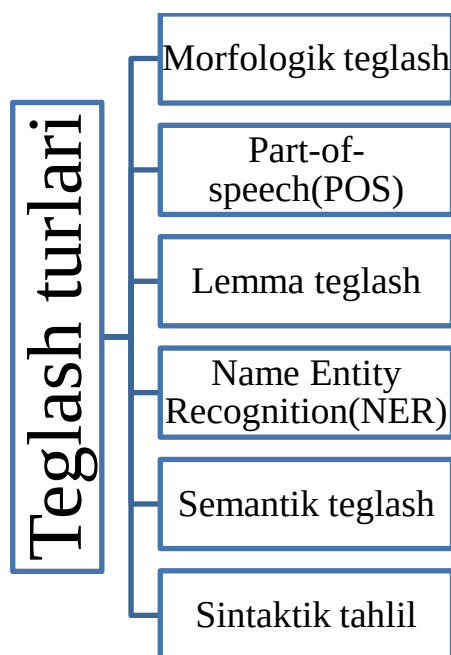
Ta’lim tizimi taraqqiy etgani sari uning tarkibida yangidan yangi sohalar ajralib chiqmoqda. Xususan, tilshunoslik sohasini oladigan bo‘lsak, ushbu fanning yangi integratsion ko‘rinishlari vujudga kelmoqda, jumladan, psixolingvistika, lingvokulturologiya, lingvoma’naviyatshunoslik kabilar shular jumlasidandir. Alqissa, kompyuter lingvistikasi tarmog‘ida shakllangan korpus lingvistikasi ham shular jumlasidandir. Korpus lingvistikasi til tarkibidagi lug‘at bazasini korpuslashtirish masalasi bilan shug‘ullanuvchi soha sanalib, bizning ushbu maqolamiz ham aynan bu yo‘nalishni qamrab oladi.

Ma’lumki, bugungi kunda tilshunoslikda korpus sohasi, ayniqsa, taraqqiy etmoqda. Korpus lingvistikasida teglash matn korpusidagi so‘zlarga grammatik, morfologik, sintaktik yoki semantik ma’lumotlarni qo‘shish jarayonini bildiradi. Bu teglash lingvistik tahlil uchun korpusdan foydalanish imkoniyatini oshiradi. Korpus lingvistikasida teglardan foydalanish vaqt o‘tishi bilan rivojlanib, texnologiya, lingvistik nazariyalar va tadqiqot metodologiyasidagi yutuqlarni aks ettirdi. Korpus lingvistikasida teglar tarixini hisoblash tilshunosligining dastlabki rivojlanishi va matn korpusida tizimli va tuzilgan lingvistik annotatsiyaga bo‘lgan ehtiyoj bilan bog‘lash mumkin [Charniak, 1997:89]. Korpus lingvistikasining dastlabki bosqichlarida tadqiqotchilar nutq qismi teglari, lemma shakllari va grammatik kategoriyalar kabi lingvistik ma’lumotlarga ega bo‘lgan matn korpusiga qo‘lda izoh

berishgan. Ushbu qo'lda izohlash jarayoni ko'p vaqt talab qildi va keng miqyosda cheklangan edi, ammo avtomatlashtirilgan teglashning keyingi ishlanmalari uchun asos yaratdi. Davrlar o'tishi bilan olib borilgan tadqiqotlar tufayli teglash jarayonining yanada taraqqiy etgan ko'rinishi yuzaga keldi.

Teglar turli tillarda amalga oshirilgan va ishlatiladigan teglar to'plami tilga qarab katta farq qiladi. Teglar, odatda, ochiq morfologik farqlarni o'z ichiga olishi uchun mo'ljallangan, ammo bu ingliz tilidagi otlar uchun emas, balki olmoshlar uchun harf belgilari kabi nomuvofiqliklarga va tillararo kattaroq farqlarga olib keladi. Yunon va lotin kabi murakkab fleksiyaga ega tillar uchun teglar miqdori juda katta bo'lishi mumkin; Inuit tillari kabi agglutinatив tillardagi so'zlarni belgilash deyarli imkonsiz. Boshqa tomondan, Petrov va boshqalar 12 toifadan iborat (masalan, otlar, fe'llar, tinish belgilari va boshqalarning pastki turlari yo'q) “universal” teglar to'plamini taklif qildilar [Charniak, 1997:56]. Juda keng teglarning juda kichik to'plami yoki aniqroq bo'lganlarning kattaroq to'plami afzalroq bo'ladimi, qo'llanilayotgan maqsadga bog'liq. Avtomatik teglash kichikroq teglar to'plamidan foydalanish jarayonini osonlashtiradi.

Korpus lingvistikasida qo'llaniladigan teglashning xilma-xil turlari mavjud. Ularni quyidagicha guruhlash mumkin [1]:



1. Part-of-speech (POS) teglash. POS teglash korpusdagi har bir so'zga grammatik kategoriya belgilashni o'z ichiga oladi. Ya'ni so'zning morfologik xususiyatlarini aniqlashda ushbu teglashdan foydalanish samarali hisoblanadi. Ushbu teglashga otlar, fe'llar, sifatlar, ergash gaplar, olmoshlar, bog'lovchilar, yuklamalar va boshqalarni teglashni kiritish mumkin. POS belgilari jumla tuzilmalarini tahlil qilish va lingvistik tadqiqotlar o'tkazishda yordam beradi.

Gap tuzilmasi	Men	Musiq	tingladim
POS teglash modeli	Pronoun(Olmosh)	Noun(Ot)	Fe'l(Verb)

Demak, keltirilgan misoldan ko‘rinib turibdiki, teglashning ushbu guruhida umumiy tarzda (morfologik va sintaktik) xususiyatlar aniqlashtiriladi va uning quyidagi teglari mavjud:

JJ – sifat (adjective)

JJR – sifatning qiyosiy darajasi (adjective, comparative)

JJS – sifatning ortirma darajasi (adjective, superlative)

MD – modal so‘zlar

NN – ot (noun)

NNS – ko‘plikdagi ot va hokazo [Wikipedia.org].

Yuqorida nomlari keltirilgan teglash usullari miqdoran ko‘p bo‘lib, teglash jarayonida so‘zning qaysi grammatik xususiyatini aks ettirishiga binoan tanlanadi.

2. Lemma teglash: Lemma teglash so‘zning asosiy yoki lug‘at shaklini aniqlashni o‘z ichiga oladi. Masalan, “yugurish” lemmasi “yugurmoq”dir. Ushbu teglash turli xil qo‘shimchalarni o‘zida jamlagan so‘zlarni tahlil qilishni soddalashtirishga yordam beradi va tadqiqotchilarga so‘zning asosiy ma’nosiga e’tibor qaratish imkonini beradi. Lemma teglashni quyidagi jadval asosida ko‘rib chiqamiz:

№	So‘zning turli shakllari	Lemma
1	Yugurish, yuguryapti, yugurajak, yugursin, yuguringlar	Yugurmoq
2	Daftari, daftarga, daftardan, daftarsiz, daftarni	Daftar
3	Qizilroq, qip-qizil, qizg‘ish, eng qizil, sal qizil	Qizil
4	Yozish, yozilajak, yozib, yozdir, yozildi, yozishmoqda	Yozmoq

3. Named Entity Recognition (NER) teglash: NER teglashi korpus ichidagi shaxslar, tashkilotlar, manzillar, sanalar va hokazo nomlari kabi nomli obyektlarni aniqlash va tasniflashni o‘z ichiga oladi. Bu teglash ma’lumot olish, matnni umumlashtirish va obyektlarni bog‘lash kabi vazifalar uchun foydalidir.

Kamil A‘zamovichning faoliyati Turkiyada 12- may kundan boshlanadi.	Shaxs(person)	Kamil A‘zamovich
	Joylashuv(Location)	Turkiya
	Sana(Date)	12-may

Yoki:

25-mayda “Tilshunoslar” guruhining tashkil topganligiga 8 yil to‘ladi	Tashkilot nomi(Organization)	“Tilshunoslar”
	Sana(Date)	25-may
	Qo‘shimcha ma’lumot(extra information)	8 yil

4. Sintaktik tahlil: Sintaktik tahlil korpusdagi gaplarning grammatik tuzilishini tahlil qilishni o‘z ichiga oladi. Bu so‘zlar va birikmalar o‘rtasidagi munosabatlarni aniqlashga yordam beradi, shu bilan birga, lingvistik tuzilmalarni sintaktik tahlil qilish imkonini beradi. Ushbu tahlil turi “Part-of-speech” tahliliy guruhiga kiritilib, uning alohida ko‘rinishi sanaladi.


Bugun havo yaxshi [Hans, 2001:69].

Keltirilgan misolda ko‘rsatilgan teglar orqali ifodalangan gapning bog‘lanish tuzilmasi aks ettirilmoqda.

5. Semantik teglash. Semantik teglash jumladagi so‘zlarning semantik rollarini aniqlashni o‘z ichiga oladi. Bu teglash jumladagi yoki so‘zlar orasidagi ma’no va munosabatlarni tushunishga yordam beradi. Ushbu teglash turi o‘z tarkibida qator guruhlarni mujassam etadi, xususan: anaforik, sifat teglash, konsept teglash, mantiqiy teglash, diskurs teglash, aspektga oid teglash. Bunda so‘zning ma’noviy xususiyati e’tiborga olinadi:

Sifat teglash	COL	<i>Colour/rang</i>	<i>Qizil, yashil, sariq, qora</i>
	QUC	<i>Quantity/miqdor</i>	<i>Ikki, yetti, o‘n besh, to‘qqiz</i>
	IST	<i>Intersective/xususiyat</i>	<i>Samimiy, yaxshi, ochiqko‘ngil</i>

6. Morfologik teglash: Morfologik teglash korpus tarkibidagi so‘zlarning morfologik belgilarini, masalan, old qo‘shimchalar, qo‘shimchalar, zamon belgilari, ko‘plik belgilari va boshqalarni tahlil qilishni o‘z ichiga oladi. Bu teglash so‘z shakllari, qo‘shimchalar va hosilalarni tushunishga yordam beradi.

Korpus tilshunosligida teglash xususida so‘z yuritar ekanmiz, uning ham o‘ziga yarasha kamchiliklari mavjud ekanligini qayd etishimiz lozim. To‘g‘ri, teglash so‘zlarni modellashtirish jarayonini osonlashtirish, biroq so‘zlarni teglash ayrim murakkablik tug‘diradi. Quyidagi misolga e’tibor bering:

Oting nima?

Qo‘lingdagi to‘pni bizga ot.

Birinchi gapda keltirilgan ot so‘zi ot so‘z turkumiga taalluqli, ikkinchi birlik esa fe‘l so‘z turkumining buyruq maylida qo‘llanmoqda. Mana shu kabi bir xil shaklli so‘zlar(omonimlar)ni korpus lingvistikasida shakllantirish teglash jarayonida birmuncha qiyinchilikka yuz tutadi. Ushbu muammoning yechimi sifatida so‘zlar bilan bog‘lanib kelgan birliklarni aniqlik bilan teglash masalasini keltirish mumkin. Sababi, birikmalarning aniq teglanishi kompyuter dasturida modellashtirish jarayonining osonlik bilan kechishiga yordam beradi.

Xulosa sifatida aytish mumkinki, ushbu teglash usullarini korpus lingvistikasida qo‘llash orqali tadqiqotchilar lingvistik tahlilning chuqurligi va ko‘lamini oshirishlari, qimmatli tushunchalarni olishlari va til ma’lumotlar to‘plamidagi tuzilmalarni aniqlashlari mumkin.

Foydalanilgan adabiyotlar:

1. Charniak, Eugene. 1997. "Statistical Techniques for Natural Language Parsing". AI Magazine 18(4):33–44.
2. Hans van Halteren, Jakub Zavrel, Walter Daelemans. 2001. Improving Accuracy in NLP Through Combination of Machine Learning Systems. Computational Linguistics. 27(2): 199–229. PDF
3. DeRose, Steven J. 1990. "Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." Ph.D. Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences. Electronic Edition available at 2.
4. D.Q. Nguyen, D.Q. Nguyen, D.D. Pham and S.B. Pham (2016). "A Robust Transformation-Based Learning Approach Using Ripple Down Rules for Part-Of-Speech Tagging." AI Communications, vol. 29, no. 3, pages 409–422.