UDK: 81'33

O‘ZBEK TILI KORPUSINI ANNOTATSIYALASH VA TEGLASH (UMUMIY QOIDA VA TIZIMLAR)

Samatboyeva Madina To‘lqinjon qizi,
tayanch doktorant
samatboyevamadina@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya: Ushbu maqolada O‘zbek tili korpuslarini tuzish jarayonida amalga oshiriladigan **annotatsiyalash** va **teglash** ishlari, ularning umumiy qoidalari hamda qo‘llanilayotgan tizimlar haqida so‘z boradi. Maqolada avvalo korpuslar uchun zarur bo‘lgan morfologik, sintaktik va semantik darajadagi teglash turlari bayon etiladi. O‘zbek tilining agglutinativ xususiyati, morfologik murakkabligi va erkin so‘z tartibi kabi xususiyatlari tufayli annotatsiyalash jarayonida yuzaga keladigan muammolar va ularni hal etish yo‘llari muhokama qilinadi.

Abstract. This article discusses the annotation and tagging processes carried out during the construction of Uzbek language corpora, their general principles, and the systems being used. The paper outlines various types of tagging required at the morphological, syntactic, and semantic levels. Due to the agglutinative nature, morphological complexity, and free word order of the Uzbek language, certain challenges arise during annotation, and the article explores potential solutions to these issues.

Аннотация: В данной статье рассматриваются процессы аннотирования и теггирования, проводимые при создании корпусов узбекского языка, их общие принципы и используемые системы. В статье описываются различные уровни теггирования: морфологический, синтаксический и семантический. Агглютинативная природа узбекского языка, его морфологическая сложность и свободный порядок слов создают определённые трудности в процессе аннотирования, и в статье обсуждаются пути их решения.

Kalit so‘zlar: *korpus, annotatsiya, teg(lash), NLP, razmetka, POS-tagging.*

KIRISH.

Korpusli lingvistika zamonaviy tilshunoslikning muhim yo‘nalishlaridan biri bo‘lib, unda tabiiy til materiallari to‘plami – korpuslar asosida lingvistik tadqiqotlar olib boriladi. Annotatsiya esa korpusdagi til birliklarini maxsus belgilar orqali tavsiflash jarayonidir. Annotatsiyalangan korpuslar tilning turli darajadagi xususiyatlarini o‘rganishda asosiy vosita bo‘lib xizmat qiladi. Ushbu maqolada

korpusli annotatsiya qilish tamoyillari, ularning nazariy asoslari, yetakchi olimlar ishlari hamda xalqaro amaliyotdagi annotatsiya tizimlari tahlil qilinadi.

So‘nggi yillarda tabiiy tilni qayta ishlash (Natural Language Processing — NLP) sohasining jadal rivojlanishi korpus lingvistikasiga bo‘lgan ehtiyojni keskin oshirdi. Ayniqsa, kompyuter lingvistikasi va sun’iy intellekt tizimlarida til birliklarini chuqur tahlil qilish zarurati korpusli annotatsiya qilish tamoyillarini ishlab chiqishga asos yaratdi. Korpusli annotatsiya — bu til birliklariga ma’lum lingvistik yoki statistik ma’lumotlarni bog‘lash, ya’ni ularni tegishli kategoriyalar bilan belgilash jarayonidir. Ushbu maqolada korpusli annotatsiya qilish tamoyillari, shu yo‘nalishda olib borilgan tadqiqot ishlari yoritiladi.

ADABIYOTLAR SHARHI.

Leech korpus annotatsiyasining umumiy tamoyillari, POS-tagging (so‘z turkumlarini belgilash) bo‘yicha asosiy standartlarni ishlab chiqqan **Geoffrey Leech** o‘zinig *“Corpus Annotation Schemes”* asarida korpus va uni tuzish tamoyillariga atroflicha to‘xtalib o‘tadi[1].

Nancy Ide va Jean Veronis o‘zining “*Introduction to the Special Issue on Corpus Annotation” (1994) asari orqali a*nnotatsiya turlari (morfologik, sintaktik va semantik) va ularning korpuslar bilan integratsiyasi haqida muhim izlanishlar olib borgan[2].

“Foundations of Statistical Natural Language Processing” asari muallifi **Christopher D. Mann** o‘z ilmiy tadqiqot ishlarida annotatsiya va statistik modellar asosida tilni avtomatik tahlil qilish metodologiyasini ishlab chiqqan[3].

ASOSIY QISM

Annotatsiya

Annotatsiyalash (lotincha *annotatio* — izoh, tushuntirish) — bu korpusdagi matnlarga lingvistik axborot qo‘shish jarayoni. Bunda matndagi har bir so‘z yoki fraza:

- so‘z turkumi (masalan, ot, fe’l, sifat – POS-tagging),
- morfologik shakllari (yig‘iq, ko‘plik, zamon),
- sintaktik roli (gapda qanday funksiya bajaradi),
- yoki hatto semantik ma’no (mazmuni, mavzusi)

bilan belgilab chiqiladi. Matnni annotatsiyalashdan maqsad – matnni kompyuterlar uchun tushunarli va tahlilga yaroqli holga keltirish. *Annotatsiya* atamasi bir vaqtning o‘zida *razmetka* atamasi bilan yonma-yon qo‘llaniladi.

Razmetka – bu matnga belgilovchi teglar (markup tags) qo‘shish jarayoni. Bu teglar lingvistik bo‘lishi ham, bo‘lmasligi ham mumkin. Misol uchun:

- HTML razmetkasi <title>, <p> kabi teglar bilan tuziladi.
- Lingvistik korpusda esa razmetka:

  [NOUN]kitob[/NOUN] [VERB]o‘qidi[/VERB]

Ko‘rinib turibdiki, razmetka *texnik belgilar orqali* ma’lumotni ko‘rsatadi.

Annotatsiya va razmetkaning **o‘xshash** tomonlari: Ikkalasi ham *matnga qo‘shimcha ma’lumot* qo‘shadi, ikkalasi ham *korpuslar bilan ishlashda* qo‘llaniladi, ikkovining maqsadi ham – *matnni strukturaviy yoki semantik jihatdan boyitish.*

Farqlari:

№	Annotatsiya	Razmetka
1	Lingvistik tahlilga qaratilgan	Ko‘proq texnik ko‘rinishdagi belgilash
2	So‘z turkumi, morfologiya, sintaksis va semantika bilan bog‘liq	Strukturaviy yoki vizual tashkil etish uchun ishlatiladi
3	Masalan: kitob → NOUN, o‘qidi → VERB	Masalan: <bold>kitob</bold>
4	Ma’lumot ko‘pincha ko‘rinmas holatda fayl tuzilmasida saqlanadi	Ma’lumot ochiq ko‘rinishda, HTML/XML kabi formatlarda bo‘ladi

Agar siz matnning tilshunoslik jihatlariga e’tibor qaratayotgan bo‘lsangiz – bu *annotatsiyalash.* Agar siz matnning strukturaviy tashkiliga yoki formatlanishiga e’tibor qaratayotgan bo‘lsangiz – bu *razmetka.*

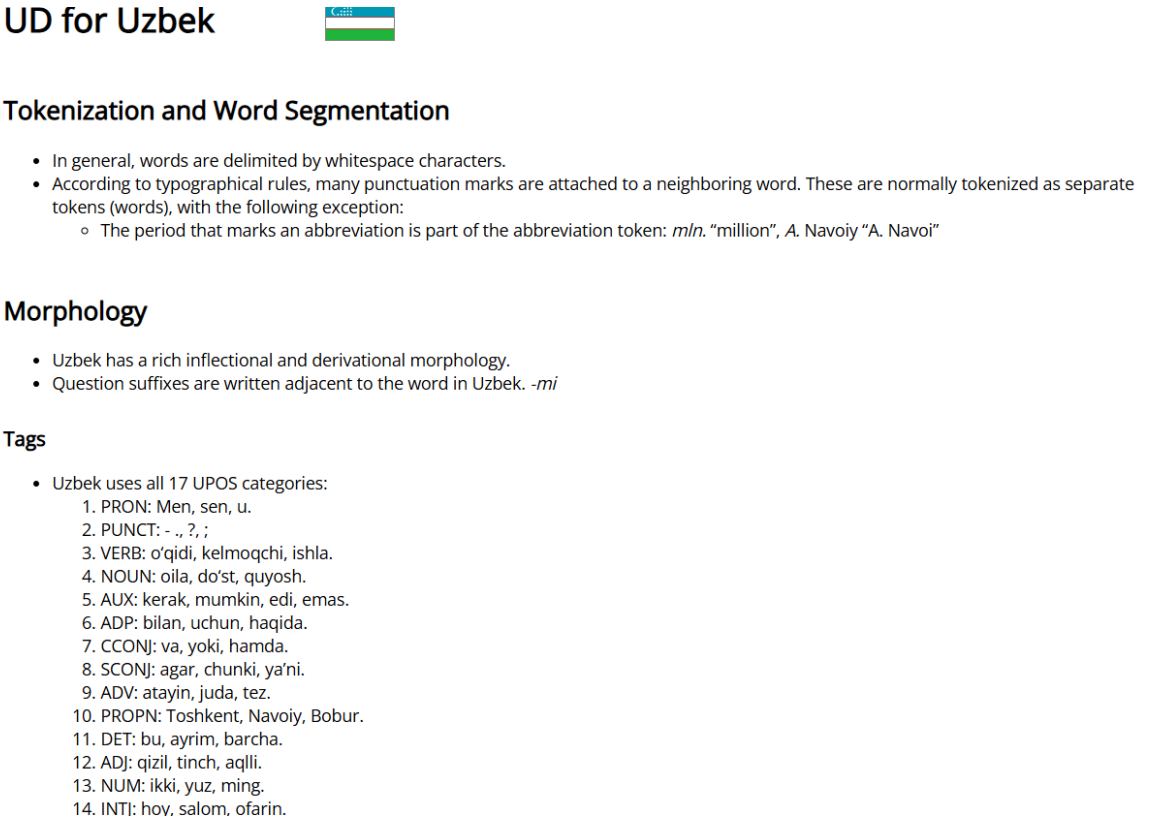
Korpus annotatsiyasi bu korpusdagi til birliklariga (so‘z, so‘z birikmalari, gap va h.k.) morfologik, sintaktik, semantik yoki pragmatik axborotlarni qo‘shish jarayonidir.

Shunday xalqaro annotatsiya korpus tizimlari mavjudki, ular juda katta hajmdagi ma’lumotlar bazasini o‘zida saqlaydi. Masalan, **“Penn Treebank (AQSh)”** (1993-yilda yaratilgan) POS-tagging va Constituency parsing (bu- gapni grammatik tuzilmasiga ko‘ra tahlil qilish usuli bo‘lib, u gapdagi so‘zlar va ularning guruhlari (ya’ni frazalar) qanday qilib bir-biriga bog‘langanini ko‘rsatadi. O‘zbekchaga tarjima qilganda**, “fraza tahlili”** yoki **“ibora tarkibiy tuzilmasini tahlil qilish”**) tizimida annotatsiya qiladi. Bu korpusda har bir so‘zga so‘z turkumi, har bir gapga esa sintaktik daraxt strukturasi biriktirilgan. Bu korpus, shuningdek, belgi (character-level) va so‘z darajasidagi (word-level) til modellari (Language Modelling) uchun ham keng qo‘llaniladi[4].

British National Corpus (BNC) - Bu ingliz tilidagi katta hajmdagi matnlar to‘plami bo‘lib, ingliz tilini o‘rganish, tadqiq qilish va tahlil qilish uchun yaratilgan. Ushbu korpus o‘z tarkibi 100 milliondan ortiq so‘zni jamlagan. POS-tagging asosida annotatsiya qilingan. Leech tomonidan ishlab chiqilgan CLAWS (Constituent Likelihood Automatic Word-tagging System) tizimi asosida annotatsiyalangan.

Yozma matnlariga: kitoblar, gazeta maqolalari, fan, san’at, din, qonun, hujjatlar va hokazolar (taxminan 90%) va og‘zaki nutq yozuvlari (spoken data): intervyular, suhbatlar, radio eshittirishlari va hokazolar (taxminan 10%) kiritilgan[5].

Shuningdek, **Universal Dependencies (UD)** (2014- hozirgi kunga qadar yangilanmoqda) – Leksik morfologik xususiyatlar, sintaktik bog‘liqliklarni tahlil qiluvchi har xil tillar uchun yagona sintaktik annotatsiya tizimidir. Ushbu tizim tarkibida 50ga yaqin tillar bazasi mavjud bo‘lib, hatto unda o‘zbek tili treebank bazasi ham kiritilgan. (Qarang: 1-rasm). O‘zbek tili treebankida 17 ta teglar ro‘yxati, morfologik tahlillar, segmentatsiya, abbreviatura, tokenlar tog‘risioda ma’lumotlar keltirilgan[6].



1-rasm. Universal Dependencies (UD)da “Uzbek treebank”i

Annotatsiya qilish quyidagi tamoyillar asosida amalga oshiriladi:

1. Aniqlik va izchillik (consistency) – har bir birlik bir xil me’yorda belgilanishi kerak.
2. Qayta foydalanish (reusability) – korpus boshqa tadqiqotlar uchun ham foydalanish imkoniyatiga ega bo‘lishi lozim.
3. Standartlashtirish (standardization) – xalqaro annotatsiya standartlariga rioya qilish kerak.

4. Ko‘p darajalilik (multi-level annotation) – korpusda morfologik, sintaktik, semantik va pragmatik annotatsiyalar mavjud bo‘lishi mumkin.

Teg(lash)

Teglar barcha turdagi transkripsiyalarni (fonetik xususiyatdan nutqiy qurilmagacha), POS teglash, ma’no belgilari, sintaktik tahlil, NER obyektlar, semantik rol belgilari, vaqt va hodisalarni aniqlash, so‘zlarning sintaktik zanjirlarini, nutq darajasini o‘z ichiga olishi mumkin [7,8,9]. Lingvistik teglashning eng muhim komponenti bu tegishli teg birligi (masalan, tovush, token yoki so‘z, birikma, fragment, hujjat) bilan bog‘lanishi kerak bo‘lgan teglar va tegishli xususiyatlarni belgilaydigan teglash sxemasi hisoblanadi. Bugungi kunda korpus matnlarini lingvistik teglashni qo‘lda yoki yarim avtomatik tarzda amalga oshirish mumkin. NLP bilan bog‘liq lingvistik bilimlarning jihatlari haqida turli xil qarashlar mavjud. Umuman olganda, NLP tadqiqotchilarining aksariyati NLP bilan bog‘liq lingvistik bilimlar, hech bo‘lmaganda, leksik, sintaktik, semantikva pragmatik xususiyatlarni o‘z ichiga olishi kerak, deb hisoblashadi [10]

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[11]

Shu o‘rinda annotatsiya va teglash atamalarining o‘xshash va farqli tomonlarini ko‘rib chiqish muhimdir. **Ikki atamaning o‘xshash tomonlari: H**ar ikkalasi ham **matnga qo‘shimcha axborot qo‘shish** bilan bog‘liq, har ikkalasi ham **korpuslarni boyitish,** ya’ni kompyuterga tilni "tushuntirish" uchun xizmat qiladi va har ikkalasi NLP (Natural Language Processing) va korpus lingvistikasi sohalarida keng qo‘llaniladi.

Farqlari:

№	Annotatsiya	Teglash
1	Kengroq tushuncha: har qanday qo‘shimcha axborotni matnga qo‘shish (POS, sintaksis, semantika, diskurs, his-tuyg‘u va h.k.)	Matndagi so‘zlarga maxsus teglar (labels) berish — ayniqsa grammatik yoki morfologik teglar (masalan: NOUN, VERB, ADJ)
2	Har xil darajadagi (morfologik, sintaktik, semantik) ma’lumotni qo‘shadi	Odatda faqat bir qatlamli axborot (ayniqsa POS-tagging) bilan chegaralanadi
3	Misol: matnda shaxs nomlarini, joy nomlarini, hissiyotlarni yoki referenslarni ajratish ham annotatsiyaga kiradi	Misol: “kitob” → NOUN, “o‘qidi” → VERB

4	Annotatsiyalashda razmetka texnologiyasi ishlatilishi mumkin (lekin bu shart emas)	Teglash — annotatsiyaning bir turi hisoblanadi

Teglash — annotatsiyaning bir ko‘rinishi. Ya’ni, har bir teglash annotatsiyadir, lekin har bir annotatsiya teglash emas.

Annotatsiya kengroq tushuncha bo‘lib, unga *teglash, semantik izoh, diskurs belgilarini qo‘shish, entitet aniqlash* kabi jarayonlar kiradi.

XULOSA

Korpus annotatsiyasi tilshunoslikda chuqur tahlil imkoniyatlarini yaratadi. Annotatsiyaning yuqorida ko‘rsatilgan tamoyillar asosida olib borilishi ilmiy tadqiqotlarning aniqligini ta’minlaydi. Yetakchi olimlar tomonidan ishlab chiqilgan tizimlar va korpuslar bu boradagi metodologik yondashuvlarni yanada takomillashtirishga xizmat qilmoqda. O‘zbek tilida ham shunday tizimli va standartlashtirilgan annotatsiyalangan korpuslar yaratish orqali milliy tilshunoslikni jahon miqyosida yangi bosqichga olib chiqish mumkin. Xulosa qilib aytganda, O‘zbek tili korpusini sifatli annotatsiyalash va teglash – tabiiy tilni qayta ishlash, avtomatik tarjima, matnni tahlil qilish kabi texnologiyalarni rivojlantirish uchun muhim bosqich hisoblanadi. O‘zbek tilining o‘ziga xos morfologik va sintaktik tuzilmalari ushbu jarayonni murakkablashtiradi, ammo moslashtirilgan teglash tizimlari va xalqaro standartlarga asoslangan yondashuvlar orqali bu muammolarni kamaytirish mumkin. Annotatsiya ishlari davomida yaratilgan korpuslar kelajakda o‘zbek tilini chuqur o‘rganish, raqamli texnologiyalarni rivojlantirish va sun’iy intellekt tizimlarida keng foydalanishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Leech, G. (1997). New Standards for Tagging and Parsing. In: Garside, R., Leech, G., McEnery, T. (Eds.) Corpus Annotation.
2. Ide, N., Veronis, J. (1994). Computational Linguistics, 19(2).
3. Manning, C., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
4. https://paperswithcode.com/dataset/penn-treebank
5. https://www.english-corpora.org/bnc/
6. https://universaldependencies.org/uz/index.html
7. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi –Toshkent, 2022. No.1. –B.154-163.
8. Bi, P. (2018). Handbook of Linguistic Annotation. Journal of Quantitative Linguistics. https://doi.org/10.1080/09296174.2018.14244955.

9. B.Boltayevich, E.B., Mirdjonovna, H.S., & Ilxomovna, A. X. (2023). Methods for Creating a Morphological Analyzer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13741 LNCS. https://doi.org/10.1007/978-3-031-27199-1_4

10. Eugene. 1997. “Statistical Techniques for Natural Language Parsing”. AI Magazine 18(4):33–44