UDK: 811.512.133`42

INCEpTION PLATFORMASI YORDAMIDA O‘ZBEK TILI UCHUN IYERARXIK BOG‘LIK DARAXTI KORPUSI YARATISH

Matlatipov San’atbek G‘ayratovich,
PhD, yetkachi ilmiy xodim
s.matlatipov@nuu.uz
O‘zbekiston Milliy universiteti

Matyoqubova Shaydo Maqsud qizi,
magistratura talabasi
shaydomatyoqubova522@gmail.com
Urganch davlat universiteti

Komilova Madinabonu Baxtiyor qizi,
magistratura talabasi
komilovamadina0920@gmail.com
Urganch davlat universiteti

Annotatsiya. Ushbu maqolada o‘zbek tilidagi matnlardan olinadigan sintaktik iyerarxik bog‘lanishli daraxt korpusini yaratish haqida so‘z boradi. Maqsad – o‘zbek tili uchun so‘zlararo sintaktik munosabatlarni aniq belgilangan daraxt ko‘rinishida tasvirlash korpusini shakllantirish. Korpus materiali sifatida zamonaviy adib Shuhrat Matkarimning “Boljon” to‘plamidan olingan 30 ta sifatli jumla tanlab olindi. Ushbu jumlalar INCEpTION platformasi yordamida ikki nafar annotator tomonidan lemmatizatsiya, morfologik belgilash va sintaktik bog‘lanishlarni qo‘lda belgilandi, so‘ngra kuratsiya (birlashtirish) bosqichi orqali yakuniy kelishuvga erishildi. Natijada, o‘zbek tilidagi hikoyaviy uslub uchun birinchi bog‘lanishli daraxt korpusi yaratildi va undagi sintaktik xususiyatlar tahlil qilindi. Asosiy natijalar shuni ko‘rsatadiki, mazkur korpus kelgusida o‘zbek tili uchun sintaktik tahlil, mashinaviy tarjima hamda ta’limiy lingvistik vositalar uchun asos bo‘la oladi.

Annotation. This article presents the development of a syntactic hierarchical dependency tree corpus based on narrative texts in the Uzbek language. The main objective is to create a corpus that systematically and clearly represents word-to-word syntactic relations in narrative-style texts of Uzbek, using a tree-based structure. As corpus material, 30 high-quality sentences were selected from “Boljon,” a short story collection by the contemporary Uzbek writer Shuhrat Matkarim. These sentences were manually annotated by two annotators using the INCEpTION platform, including lemmatization, morphological tagging, and

syntactic dependency annotation. The final version of the annotations was produced through a curation phase, achieving consensus between annotators. As a result, the first syntactic dependency tree corpus for Uzbek narrative style was created and syntactic features within the corpus were analyzed. The findings demonstrate that this corpus can serve as a valuable foundation for syntactic parsing, machine translation, natural language processing, and educational linguistic tools for the Uzbek language.

Аннотация. В данной статье рассматривается создание синтаксического иерархического корпуса с древовидной структурой на основе повествовательных текстов на узбекском языке. Основная цель исследования — сформировать корпус, в котором синтаксические связи между словами в текстах повествовательного жанра узбекского языка будут чётко и систематично представлены в виде древовидной структуры. В качестве материала корпуса были отобраны 30 качественных предложений из сборника рассказов «Больжон» современного узбекского писателя Шухрата Маткарима. Эти предложения были вручную аннотированы двумя аннотаторами с использованием платформы INCEpTION, включая лемматизацию, морфологическую разметку и синтаксические зависимости. Финальная версия корпуса была сформирована в результате этапа кураторской сверки и согласования аннотаций. В результате был создан первый синтаксический корпус с древовидными зависимостями для повествовательного стиля узбекского языка, в котором были проанализированы синтаксические особенности. Полученные результаты показывают, что данный корпус может стать важной основой для синтаксического анализа, машинного перевода, обработки естественного языка и разработки образовательных лингвистических ресурсов для узбекского языка.

Kalit so‘zlar: *o‘zbek tili, hikoyaviy matnlar, sintaktik bog‘lanish, daraxt korpusi, lemmatizatsiya, morfologik belgilash, Universal Dependencies, INCEpTION platformasi, annotatsiya, kompyuter lingvistikasi, tabiiy tilni qayta ishlash (NLP)*

O‘zbek tili Markaziy Osiyodagi eng yirik tillardan biri bo‘lib, so‘zlashuvchilari soni 30–40 million atrofida baholanadi. Turk tillari orasida turk tilidan keyin ikkinchi eng ko‘p so‘zlashuvchiga ega til hisoblanishiga qaramay, o‘zbek tili uchun kompyuter lingvistikasi (NLP) sohasida resurslar juda cheklangan. Xususan, matnlarni sintaktik daraxt shaklida belgilash uchun treebank (daraxt

korpusi)lar deyarli mavjud emas edi. Bunday korpuslar esa bugungi kunda ko‘plab tillar uchun yaratilgan bo‘lib, ular orqali turli NLP modellari o‘qitiladi va lingvistik tadqiqotlar olib boriladi.

Ayrim so‘nggi tadqiqotlarda o‘zbek tilida turli korpus va modellarning paydo bo‘layotgani kuzatilmoqda: masalan, UZWORDNET leksik-semantik tarmog‘i yaratildi, matnlarni tuyg‘u (sentiment) yorliqlari bilan belgilangan to‘plamlar va matn tasnifi korpuslari taklif qilindi (Kuriyozov va boshq., 2019; Rabbimov va Kobilov, 2020), hatto o‘zbek tili uchun UzBERT hamda BERTbek kabi yirik transformator modellar dastlabki o‘qitish orqali tayyorlandi. Shunga qaramay, hali ham o‘zbek tilida to‘liq qo‘lda belgilangan sintaktik daraxt korpusi deyarli yo‘q edi. 2025-yilda o‘zbek tili uchun ilk Universal Dependencies (UD) daraxt korpusi taqdim etildi, u 500 ta jumlani (taxminan 5850 token) qamrab oladi Biroq, mazkur UD korpusi yangiliklar va badiiy matnlardan avtomatik va qo‘lda belgilash aralash usulida tuzilgan). O‘zbek tilidagi hikoyaviy (badiiy) uslub xususiyatlarini to‘la aks ettiruvchi, mutlaqo qo‘lda, yuqori sifatli sintaktik korpus hali yaratilmagan edi.

Ushbu loyiha ana shu bo‘shliqni to‘ldirishga qaratilgan. Biz o‘zbek tilidagi hikoyaviy matnlardan kichik hajmli bo‘lsa-da, sifatli va murakkab sintaktik bog‘lanishlarni o‘zida jamlagan daraxt korpusini yaratdik. Korpus tuzishda Universal Dependencies tamoyillariga asoslandik, ya’ni so‘z turkumlari va grammatik kategoriyalar xalqaro standart (UD) bo‘yicha belgilandi. Natijada o‘zbek tilini kompyuterda qayta ishlash (NLP) uchun yangi bir resurs – hikoyaviy matnlarga mo‘ljallangan treebank hosil qilindi. Quyida ushbu korpusni shakllantirish metodologiyasi, uni belgilash jarayoni, annotatorlararo kelishuv ko‘rsatkichlari hamda dastlabki tahliliy natijalar yoritiladi.

Adabiyotlar tahlili va metodologiya

O‘zbek tilini kompyuter lingvistikasi yo‘nalishida qayta ishlash bo‘yicha dastlabki ishlar morfologik tahlil va so‘z turkumlarini belgilashdan boshlangan. Masalan, Matlatipov va Vetulani (2009) Prolog muhiti yordamida o‘zbek tilining morfologik tahlilchisini yaratgan, biroq ushbu tizim murakkab so‘z formalarini to‘liq qo‘llab-quvvatlay olmagan. Keyinchalik o‘zbek tilining grammatik xususiyatlarini chuqurroq aks ettiruvchi qo‘shimcha belgilar taklif qilindi (Sharipov va boshq., 2022) va UzbekTagger deb nomlangan qoidaviy qatorga asoslangan so‘z turkumi teglagichi ishlab chiqildi. Ushbu POS-tegger o‘zbek tilida 12 ta asosiy so‘z

turkumini aniqlab, qo‘lda belgilangan kichik korpusda sinovdan o‘tkazilgan (Sharipov va boshq., 2023). Morfologik tahlilda so‘zning negizini (stem) ajratish va lemmalash bo‘yicha ham ishlanmalar mavjud: UzbekStemmer (Sharipov & Yuldashev, 2022) algoritmi va so‘ngra UzMorphAnalyzer tizimi (Salaev, 2023) taqdim etilgan. UzMorphAnalyzer o‘zbek tilida so‘zlarni tarkibiy qismlarga ajratib, ularning lug‘aviy ma’nosini aniqlash (stemmer+lemmatizer+morfologik analiz) imkonini beradi. Shuningdek, UZWORDNET (Agostini va boshq., 2021) singari leksik ma’lumot bazalari ham yaratilgan bo‘lib, o‘zbek tilida sinonimik bog‘lanishlar tarmog‘ini taklif qiladi. Biroq, tilning sintaktik tuzilishini aks ettiruvchi bog‘lanishli daraxt korpusi masalasi e’tibordan chetda qolib kelgan edi. Yaqinda Akhundjanova va Talamo (2025) ilk bor o‘zbek tilining Universal Dependencies treebankini taqdim etdilar. [1] UD_Uzbek-UT deb nomlangan ushbu korpus 500 ta jumla (5850 token)dan iborat bo‘lib, yangiliklar va badiiy adabiyotlardan olingan jumlalar avtomatik analiz va qo‘lda tuzatish kombinatsiyasi orqali belgilangan. Bizning ishimiz esa to‘liq qo‘lda va faqat hikoyaviy uslubga tegishli jumlalarni belgilash orqali o‘zbek tilidagi sintaktik bog‘lanishlarni batafsilroq ochib berishga qaratilgan.[2]

Korpus uchun ma’lumot manbasi sifatida adib Shuhrat Matkarimning “Boljon” nomli hikoyalar to‘plami tanlandi. Ushbu to‘plamdagi matnlardan 30 ta murakkab, grammatik jihatdan boy jumlalar qo‘lda ajratib olindi. Jumlalarning umumiy uzunligi va mazmuniy xilma-xilligi e’tiborga olindi – dialog va monologlar, sodda va qo‘shma gaplar, turli sintaktik konstruktsiyalar kiritildi. Shundan so‘ng har bir jumla tokenlarga ajratildi (so‘z va alomatlarga bo‘linib, maxsus ID raqamlar bilan tartiblandi), har bir token uchun lemma (asosiy lug‘aviy shakl) belgilandi, so‘z turkumi (masalan, ot, fe’l, sifat va hokazo) va uning morfologik xususiyatlari (masalan, son, kelishik, zamon, kishilik) aniqlanib teglar qo‘yildi. So‘ngra jumla tarkibida har bir so‘zning grammatik jihatdan qaysi so‘zga bog‘lanishi (kimga ega, kimni to‘ldiruvchi, qaysi so‘zga aniqlovchi va hokazo) bog‘lanish relatsiyasi ko‘rinishida ko‘rsatildi. Barcha belgilash ishlari INCEpTION dasturiy platformasi yordamida amalga oshirildi. Ikki nafar mustaqil annotator har bir jumlani mustaqil belgiladi. Annotatsiya jarayoni quyidagi bosqichlardan iborat bo‘ldi:

1. Lemmatizatsiya va so‘z turkumini belgilash: Annotatorlar jumladagi har bir so‘z uchun uning boshlang‘ich lemmasini va Universal Dependencies tagiga mos so‘z turkumini belgilandi (masalan, otlar uchun NOUN, fe’llar uchun VERB, ravishlar uchun ADV va hokazo).

2. Morfologik kategoriyalarni belgilash: Har bir tokenning grammatik kategoriyalari aniqlanib yozildi. Masalan, otlarda kelishik va son (Case, Number xususiyatlari), fe’llarda zamon, nisbat, shaxs va son (Tense, Voice, Person, Number) kabi attributlar belgilandi. Bu belgilar UD standartidagi qisqacha taglar yordamida ifodalandi.

3. Sintaktik bog‘lanishlarni aniqlash: Jumladagi har bir so‘zning sintaktik jihatdan qaysi so‘zga va qanday munosabat bilan bog‘langanligi ko‘rsatildi. Bunda har bir bog‘lanish turi UD relatsiyalar inventaridan tanlandi (masalan, ega uchun nsubj, to‘ldiruvchi uchun obj, aniqlovchi uchun amod, hol uchun obl, kesim uchun root va hokazo). Har bir token tegishli “bosh so‘z” (head)ning ID raqami bilan bog‘landi va bog‘lanish turi yozildi. Masalan, kesim fe’l odatda gapning “root”i bo‘lib, boshqa so‘zlar bevosita yoki bilvosita unga bog‘lanadi.

Ikki annotator mustaqil belgilab chiqqach, kuratsiya bosqichi o‘tkazildi. INCEpTION platformasining curation rejimida annotatorlar kiritgan belgilardagi tafovutlar ko‘rishli tarzda taqqoslanib, yakuniy kelishilgan variant shakllantirildi. Bu bosqichda har bir zid holat muhokama qilinib, umumiy to‘g‘ri belgi tanlandi. Kuratsiya natijasida bir xil ID raqamli jumlalar uchun yakuniy “oltin standart” annotatsiya olindi.

Eksperimentlar va natijalar

Yaratilgan korpus kichik hajmli bo‘lsa-da, uning tarkibiy statistikasi va xususiyatlari tahlil qilindi. Korpus quyidagi ko‘rsatkichlarni o‘z ichiga oladi:

- Jumla soni: 30 ta
- Tokenlar soni (umumiy): ~600 ta
- O‘rtacha jumla uzunligi: 20 ta token atrofida
- Turli (unikal) leksemalar: ~400 ta

Yuqoridagi taxminiy taqsimotdan ko‘rinadiki, hikoyaviy matnlarda otlar va fe’llar eng yuqori ulushga ega, bu esa tabiiy holdir – matn asosan voqea-hodisalarni (fe’l orqali) va ishtirokchilarni (ot orqali) ifodalaydi. Korpusdagi har bir jumla sintaktik daraxt ko‘rinishida annotatsiya qilingan. Quyida korpusdan olingan bir misolni CoNLL-U formatida keltiramiz (har bir token uchun ID, so‘zning o‘zi, lemma, asosiy turkum, morfologik grammema, sintaktik bosh ID va bog‘lanish turi ko‘rsatilgan):

1 Boljon Boljon PROPN _ Case=Nom|Number=Sing 5 nsubj _ _

2 kechasi kecha ADV _ _ 5 advmod _ _

3 bog‘da bog‘ NOUN _ Case=Loc|Number=Sing 5 obl _ _

4 sayr sayr NOUN _ Case=Nom|Number=Sing 5 obj _ _

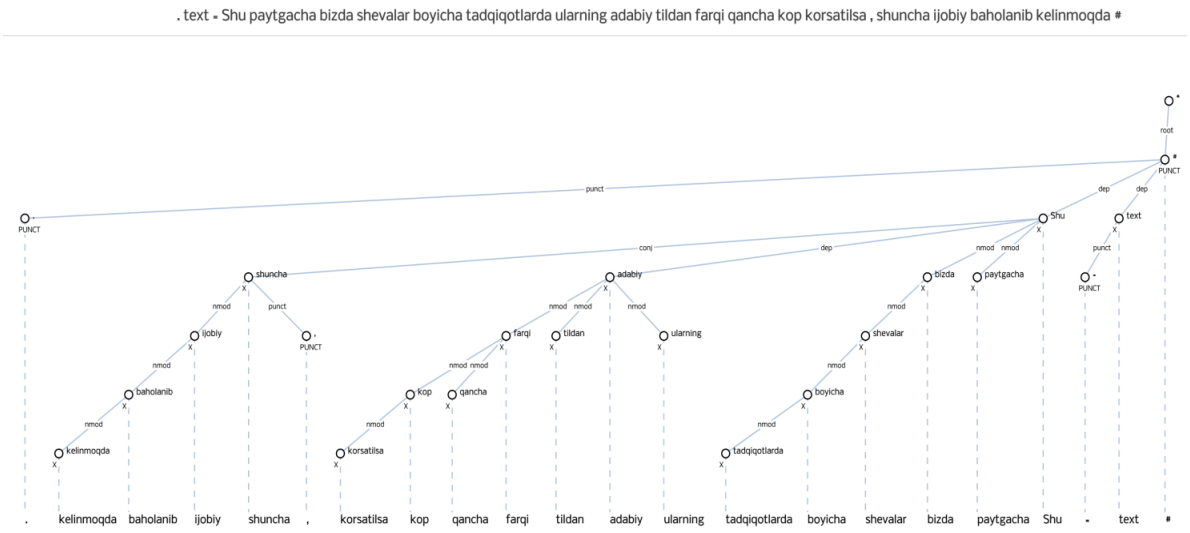
5 qildi qil VERB _ Tense=Past|Person=3|Number=Sing 0 root _ _

6 . . PUNCT _ _ 5 punct _ _

Misol: Boljon kechasi bog‘da uni ko‘rdi. Ushbu jumlada "ko‘rdi" fe’li kesim bo‘lib, root sifatida belgilangan. Boshqa so‘zlar bevosita unga bog‘langan: Boljon – nsubj (ega, kim ko‘rdi?), kechasi – advmod (qachon ko‘rdi?), bog‘da – obl (qayerda ko‘rdi?), uni – obj (kimni ko‘rdi?) munosabatlari bilan bog‘langan. Bu orqali, jumladagi har bir so‘zning grammatik vazifasi va gap tuzilishidagi o‘rni daraxt shaklida ko‘rinadi.[3]

Korpus tahlili jarayonida o‘zbek tilining hikoyaviy uslubiga xos bir qator qiziqarli sintaktik va morfologik holatlar kuzatildi. Avvalo, egasiz gaplar (subyekti tushirilgan, shaxs-son grammatik ko‘rsatkichlari fe’lda mavjud) uchrashi kuzatildi. Masalan, ba’zi jumlalarda gapning egasi yozilmagan bo‘lsa-da, fe‘ldagi shaxs va son qo‘shimchasidan (3-shaxs birlik) uning “u” (3-shaxs) ekani anglashiladi. Bunday holatlar o‘zbek tilining null-subject (pro-drop) xususiyatini tasdiqlaydi va korpusimizda bunday gaplar aniq nsubj:implic yoki shunga o‘xshash usulda belgilandi.

Shuningdek, bog‘langan va ravishdosh vositali qo‘shma gaplar ko‘pligi e‘tiborga molik bo‘ldi. Hikoyaviy jumlalarda bir nechta sodda gaplar va, ham kabi bog‘lovchilar yoki -sa kabi qo‘shimchalar orqali bog‘lanib, murakkab gaplar hosil qilingan. Masalan: “U hovliga chiqib, atrofga nazar tashladi va birpas sukut saqladi.” kabi jumlalar bir vaqtning o‘zida ikkita-yu uchta sodda harakatni ifodalaydi. Korpusimizda bunday hollarda ravishdoshli fe’lning bog‘lanishi advcl (hol ergash gapli qo‘shma gap) yoki conj (bog‘lovchisiz qo‘shma gap) sifatida belgilandi, kontekstga qarab. Natijada, daraxt tuzilishida bir necha rootga ega bo‘lmagan, balki bitta bosh fe’l atrofiga birikkan struktura hosil bo‘ladi (ketma-ket bog‘langan fe’llar).Bu o‘zbek tilining polisentetik sintaksisini ko‘rsatadi.(1-rasm)



1-rasm. Sintaktik bog‘lanishli bog‘lik daraxt

Morfologik jihatdan qaralganda, otlarda kelishik kategoriyasining faol qo‘llanilishi kuzatildi: korpusdagi otlarning ~60% atrofida turli kelishik qo‘shimchalari mavjud (-ni, -ga, -da, -dan va hokazo). Bu hol, o‘zbek tilida predloglarning yo‘qligi va ularning vazifasini asosan qo‘shimchalar bajarishini tasdiqlaydi. Masalan, yuqoridagi misolda bog‘da so‘ziga locative kelishik qo‘shimchasi -da qo‘shilib, ingliz tilidagi “in the garden” ma’nosini bitta so‘z orqali ifodalagan. Shuningdek, ba’zi hollarda tushum kelishigi (akuzativ) qo‘shimchasi tushirib qoldirilgan (vositasiz to‘ldiruvchi) holatlar ham uchradi, bu esa semantik noaniqlikni keltirib chiqarmagan.

Yuqoridagi kuzatishlar bizning korpusimizni rus va ingliz tillaridagi treebanklar bilan solishtirganda qiziqarli farqlarni ko‘rsatadi. Masalan, ingliz tilida ega yoki to‘ldiruvchi tushirilishi deyarli kuzatilmaydi, so‘z tartibi qattiqroq, kelishik degan kategoriya yo‘q. Rus tilida esa kelishiklar mavjud, ammo unda otlarda grammatik rod (jins) kategoriyasi bor, o‘zbek tilida esa jins kategoriyasi yo‘q. Shuningdek, rus tilida ham egani tushirish mumkin (masalan, buyruq gaplarda), lekin o‘zbek tilidagidek harakat fe’l ma’noviy shaxsini qo‘shimchadan bilish odatiy emas (rus tilida fe’l shakli subyektga uncha moslashmaydi 3-shaxsdan tashqari). Turk tiliga solishtirilsa, o‘zbek tilining sintaksisi ko‘p jihatdan o‘xshash: so‘z tartibi erkin SOV, so‘zlar agglutinativ quriladi, artikl yo‘q. Biroq, o‘zbek tilida tarixiy fonetik o‘zgarishlar tufayli turk tilidagidek kuchli tovush moslashuvi (vowel harmony) emas, bir xil qo‘shimchalar qo‘llanadi, bu esa morfologik tahlilni biroz soddalashtirishi mumkin.

INCEpTION platformasida ishlash jarayoni ancha qulay bo‘lib, unda foydalanuvchilar aniq interfeys orqali tokenlarni belgilash, relatsiyalar o‘rnatish va tahrir qilish imkoniyatiga ega. Annotatsiya turlarining ko‘rsatilgan rangi, bog‘lanish yo‘nalishlari va o‘zaro aloqador tokenlar interaktiv ko‘rinishda tasvirlanadi. Shu bilan birga, platforma annotatorlararo kelishuv (inter-annotator agreement)ni baholash uchun avtomatik vositalarni ham taklif qiladi. Bu vositalar yordamida lemmatizatsiya, POS teglash va dependency relatsiyalardagi farqlar hisoblab chiqiladi va kappa koeffitsienti kabi statistik ko‘rsatkichlar orqali umumiy moslik darajasi aniqlanadi.

Xulosa o‘rnida ta’kidlash joizki, mazkur ish natijasida o‘zbek tilining hikoyaviy uslubi uchun dastlabki kichik hajmdagi, ammo boy sintaktik axborotga ega bo‘lgan bog‘lanishli daraxt korpusi yaratildi. Ushbu korpusning asosiy ahamiyati shundaki, u o‘zbek tilida sintaktik parsing (avtomatik tahlil) masalalari uchun qo‘lda belgilangan namunaviy ma’lumotlar to‘plamini taqdim etadi. Korpusdan foydalangan holda, kelgusida o‘zbek tili uchun stochastic parsers (masalan, neyron tarmoq asosidagi DEP parserlar)ni o‘qitish mumkin bo‘ladi. Shuningdek, korpus mashinaviy tarjima sohasida, xususan, o‘zbek tilining sintaktik tuzilishini model qilishda asqotadi – tarjima tizimlari bu korpusdan o‘zbekcha gap tuzilishini o‘rganishi, boshqa tillar daraxt korpusi bilan taqqoslab, mos keluvchi strukturalarni topishi mumkin. Korpusdan o‘rganilgan qoidalar va andozalar ta’limiy dasturlarda, masalan, o‘zbek tili grammatikasini o‘rgatuvchi interaktiv ilovalarda qo‘llanilishi mumkin: o‘quvchilar jumlalarni tahlil qilishni aynan shu korpus misollarida mashq qilishlari mumkin.

Foydalanilgan adabiyotlar:

1. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021). UZWORDNET: O‘zbek tili uchun leksik-semantik ma’lumotlar bazasi. 11-Global WordNet konferensiyasi materiallari, sah. 8–19, Janubiy Afrika (UNISA).

2. Akhundjanova, A., & Talamo, L. (2025). O‘zbek tili uchun Universal Dependencies daraxt korpusi. RESOURCEFUL-2025: Qo‘llab-quvvatlanishi past tillar va domenlar uchun resurslar bo‘yicha 3-Workshop materiallari. ACL..

3. Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodríguez, C. (2019). “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek.” In *Human Language Technologies as a Challenge for Computer Science and Linguistics: 9th Language and Technology Conference (LTC 2019)*, Poznan, Poland. Revised Selected Papers, pp. 232–243. Berlin: Springer.

4. Kuriyozov, E., Vilares, D., & Gómez-Rodríguez, C. (2024). BERTbek: o‘zbek tili uchun oldindan o‘qitilgan transformator modeli. 3-SIGUL (LREC-COLING 2024) anjumani materiallari, sah. 33–44.

5. Mansurov, B., & Mansurov, A. (2021). UzBERT: o‘zbek tilida BERT modelini dastlabki o‘qitish. *arXiv preprint* arXiv:2108.09814.

6. Matlatipov, G., & Vetulani, Z. (2009). Prolog yordamida o‘zbek tili morfologik tahlili. In *Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Turkic Languages*.

7. Rabbimov, I. M., & Kobilov, S. S. (2020). Multi-class text classification of Uzbek news articles using machine learning. *Journal of Physics: Conference Series*, 1546(1), 012097.

8. Rahmatullayev, Sh. (2006). *Hozirgi adabiy o‘zbek tili* (darslik). Toshkent: Universitet nashri.

9. Salaev, U. (2023). UzMorphAnalyzer: O‘zbek tili uchun morfologik tahlil modeli (so‘zlardagi qo‘shimchalarga asoslangan). *Science and Innovation*, 2(1), 29–34.

10. Sharipov, M., Mattiyev, J., Sobirov, J., & Baltayev, R. (2022). O‘zbek tilining morfologik va sintaktik teglangan korpusi-ni yaratish. ALTNLP 2022: *Agglutinative Language Technologies as a Challenge of NLP* (virtual konferensiya, Koper, Sloveniya), CEUR-WS, vol. 3315, sah. 93–98.

11. Sharipov, M., & Sobirov, O. (2022). O‘zbek tili uchun qoidaviy lemmalash algoritmini ishlab chiqish. ALTNLP 2022 konferensiyasi materiallari, CEUR-WS, vol. 3315, sah. 154–159.

12. Sharipov, M., & Yuldashev, O. (2022). UzbekStemmer: O‘zbek tili uchun qoidaviy stemming algoritmi. ALTNLP 2022 konferensiyasi materiallari, CEUR-WS, vol. 3315, sah. 137–144.

13. Sharipov, M., Kuriyozov, E., Yuldashev, O., & Sobirov, O. (2023). UzbekTagger: O‘zbek tili uchun qoidaviy so‘z turkumlari teglagichi. *arXiv preprint* arXiv:2301.12711.

14. Shuhrat Matkarim.(2017). Kun qanday boshlanadi. – G‘afur G‘ulom nashriyoti.

15. Universal Dependencies Treebank for Uzbek .

16. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation - ACL Anthology