



UDK: 811.512(575.1)

San'atbek MATLATIPOV,
O'zbekiston Milliy universiteti yetakchi ilmiy xodimi, PhD
E-mail: s.matlatip@nuu.uz
Shaydo MATYOQUBOVA,
Urganch davlat universiteti magistranti
Madinabonu KOMILOVA,
Urganch davlat universiteti magistranti

O'zMU professori N.Sa'dullayeva taqrizi asosida

A HIERARCHICALLY DEPENDENCY LINKED SYNTACTIC TREEBANK FOR UZBEK LITERARY TEXTS Annotation

This paper examines the establishment of a syntactic hierarchical treebank derived from Uzbek literary works, with the objective of advancing this corpus. Thirty high-quality sentences were extracted from "Boljon," a short tale anthology by modern Uzbek author Shuhrat Matkarim, serving as corpus material. These sentences were manually annotated by two annotators using the INCEpTION platform, including lemmatization, morphological tagging, and syntactic dependency annotation. As a result, the first syntactic dependency tree corpus for Uzbek narrative style was created and syntactic features within the corpus were analyzed.

Key words: Uzbek language, literary texts, syntactic connection, tree corpus, lemmatization, morphological marking, Universal Dependencies, INCEpTION platform, annotation.

ИЕРАРХИЧЕСКИЙ СВЯЗАННЫЙ СИНТАКСИЧЕСКИЙ КОРПУС ДЛЯ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ НА УЗБЕКСКОМ ЯЗЫКЕ

Аннотация

В данной статье рассматривается вопрос создания и совершенствования корпуса синтаксических иерархических деревьев на основе узбекских литературных произведений. Был разработан корпус, состоящий из 30 высококачественных предложений, отобранных из сборника повестей современного узбекского писателя Шухрата Маткарима. Эти предложения были аннотированы двумя экспертами-специалистами с использованием платформы INCEpTION, включая лемматизацию, морфологическую разметку и определение синтаксических связей. В результате впервые был создан корпус синтаксических иерархических деревьев для узбекского повествовательного стиля и проанализированы его синтаксические особенности.

Ключевые слова: Узбекский язык, художественные тексты, синтаксическая связь, древовидный корпус, лемматизация, морфологическая разметка, универсальные зависимости, платформа INCEpTION, аннотация.

O'ZBEK TILI BADIY MATNLARI UCHUN IYERARXIK BOG'LIQLIK DARAXTI KORPUSI

Аннотация

Ushbu maqolada o'zbek adabiy asarlaridan olingan sintaktik iyerarxik daraxt korpusini yaratish va uni takomillashtirish masalasi o'rganiladi. Zamonaviy o'zbek yozuvchisi Shuhrat Matkarimning qissalar to'plamidan saralab olingan 30 ta yuqori sifatli jumladan tashkil topgan korpus ishlab chiqildi. Bu jumlar INCEpTION platformasidan foydalanib, ikki ekspert mutaxassis tomonidan annotatsiyalangan bo'lib, lemmatizatsiya, morfologik teglash va sintaktik bog'liqliklarni aniqlashlarni o'z ichiga oladi. Natijada, o'zbek hikoya uslubi uchun ilk bor sintaktik iyerarxik daraxti korpusi yaratildi va undagi sintaktik xususiyatlar tahlil qilindi.

Kalit so'zlar: O'zbek tili, badiiy matnlar, sintaktik bog'lanish, daraxt korpusi, lemmatizatsiya, morfologik belgilash, Universal Dependencies, INCEpTION platformasi, annotatsiya.

Kirish. 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[1], 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[2]. 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.

Ushbu maqolada 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.

Mavzuga oid adabiyotlarning tahlili. 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 Prolog muhiti yordamida o'zbek tilining morfologik tahlilchisini yaratgan, biroq ushbu tizim murakkab so'z formalarini to'liq qo'llab-quvvatlay olmagan[3]. Keyinchalik o'zbek tilining grammatik xususiyatlarini chuqurroq aks ettiruvchi qo'shimcha belgilar taklif qilindi[4]. Morfologik tahlilda so'zning negizini (stem) ajratish va lemmalash bo'yicha ham ishlamlar mavjud: UzbekStemmer[5] algoritmi va so'ngra UzMorphAnalyzer tizimi[6] taqdim etilgan. UzMorphAnalyzer o'zbek tilida so'zlarni tarkibiy qismlarga ajratib, ularning lug'aviy ma'nosini aniqlash (stemmer+lemmatizer+morfologik analiz) imkonini beradi. Biroq, tilning sintaktik tuzilishini aks ettiruvchi bog'lanishli daraxt korpusi masalasi e'tibordan chetda qolib kelgan edi. Bizning ishimiz esa to'liq qo'lda va faqat hikoyaviy uslubga tegishli jummalarni belgilash orqali o'zbek tilidagi sintaktik bog'lanishlarni batafsilroq ochib berishga qaratilgan.

Tadqiqot metodologiyasi. Korpus uchun ma'lumot manbasi sifatida adib Shuhrat Matkarimning "Boljon" nomli hikoyalar to'plami tanlandi[7]. Ushbu to'plamdagi matnlardan 30 ta murakkab, grammatik jihatdan boy jumlar qo'lda ajratib

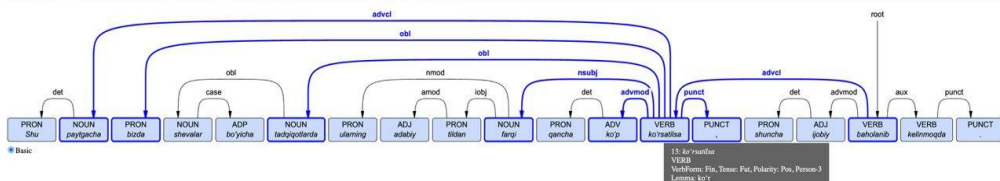
olindi. Jumalarning umumiy uzunligi va mazmuniy xilma-xilligi e'tiborga olindi – dialog va monologlar, sodda va qo'shma gaplar, turli sintaktik konstruksiyalar kiritildi. Shundan so'ng har bir jumla tokenlarga ajratildi (so'z va alomatlar bo'linib, maxsus ID raqamlar bilan tartiblandi), har bir token uchun lemma (asosiy lug'aviy shakl) belgilandi, so'z turkumi va uning morfologik xususiyatlari 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 ekspert annotator har bir jumlaning 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.

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 atributlar 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.(1-rasm).

text = Shu paytgacha bizda shevalar bo'yicha tadqiqotlarda ularning adabiy tildan farqi qancha ko'p ko'rsatilsa, shuncha ijobiy baholanib kelinmoqda.



1-rasm. Sintaktik bog'lanishli daraxt

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 jumlar uchun yakuniy "oltin standart" annotatsiya olindi.

Annotatorlararo kelishuv darajasini o'lchash uchun ikki xil ko'rsatkich — Krippendorff alfasini (α) va Cohen kappasi (κ) hisoblab chiqildi. Krippendorff alfasini annotatorlar o'rtasidagi tasodifiy moslik ehtimolini inobatga olgan holda umumiy kelishuvni ko'rsatadi;

Bizning korpus belgilash jarayonimizda juda yuqori annotatorlararo kelishuvga erishildi. Xususan:

Lemmatizatsiya va so'z turkumlari bo'yicha: $\alpha = 0.95$; $\kappa = 0.97$ — deyarli mukammal kelishuv.

Morfologik xususiyatlar bo'yicha: $\alpha = 0.77$; $\kappa = 0.78$ — yuqori darajadagi kelishuv.

Sintaktik bog'lanishlar bo'yicha: $\alpha = 0.81$; $\kappa = 0.82$ — yuqori darajadagi kelishuv.

Umuman olganda, κ va α qiymatlari 0.8 atrofida yoki undan yuqori bo'lganda annotatsiya ishonchli hisoblanadi. Demak, bizning belgilash jarayonimizda annotatorlar orasida

juda yaxshi uyg'unlik mavjud bo'lib, bu korpusning sifatli ekanidan dalolat beradi.

Tahlil 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

So'z turkumlari tarkibi: Ot (NOUN) – ~30%, Fe'l (VERB) – ~15%, Sifat (ADJ) – ~10%, Olmosh (PRON) – ~10%, Ravish (ADV) – ~5%, Yuklama/Ko'makchi (ADP/AUX) – ~5%, Bog'lovchi (CCONJ/SCONJ) – ~5%, Boshqalar (NUM, INTJ, etc.) – ~5%, Punktatsiya – ~15%.

Yuqoridagi taxminiy taqsimotdan ko'rinadiki, badiiy 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-jadval):

# text = Shu paytgacha bizda shevalar bo'yicha tadqiqotlarda ularning adabiy tildan farqi qancha ko'p ko'rsatilsa, shuncha ijobiy baholanib kelinmoqda.					
No	So'z	Lemma	POS	Morphological features	Dependency

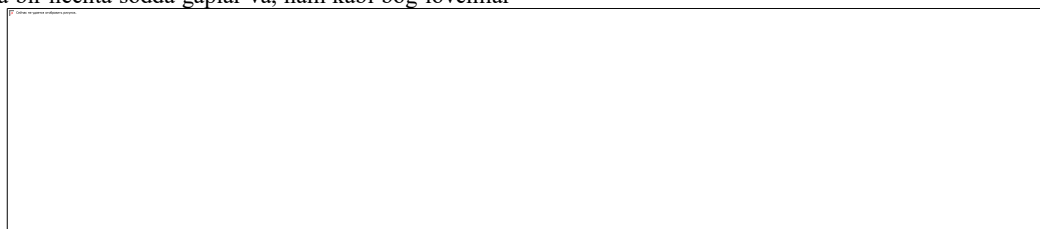
1	Shu	shu	PRON	PronType=Dem Number=Sing Case=Nom	Det
2	paytgacha	payt	NOUN	Number=Sing Case=Nom	Advcl
3	bizda	biz	PRON	PronType=Prs Number=Plur Case=Loc	Obl
4	shevalar	sheva	NOUN	Number=Plur Case=Num	Obl
5	bo'yicha	bo'yicha	ADP	-	Case
6	tadqiqotlarda	tadqiqot	NOUN	Number=Plur Case=Loc	Obl
7	ularning	u	PRON	PronType=Dem Number=Plur Case=Gen	Nmod
8	adabiy	adabiy	ADJ	Degree=Pos	Amod
9	tildan	til	PRON	Number=Sing Case=Abl	Iobj
10	farqi	farq	NOUN	Number=Sing Case=Nom	Nsubj
11	qancha	qancha	PRON	PronType=Int Number=Sing Case=Nom	Det
12	ko'p	ko'p	ADV	-	Advmod
13	ko'rsatilsa	ko'r	VERB	VerbForm=Fin Tense=Fut Polarity=Pos Person=3	Advcl
14	.	-	PUNCT	-	Punct
15	shuncha	shu	PRON	PronType=Dem Number=Sing Case=Nom	Det
16	ijobiy	ijobiy	ADJ	Degree=Pos	Advmod
17	baholanib	bahola	VERB	VerbForm=Conv Polarity=Pos	Root
18	kelinmoqda	kel	VERB	VerbForm=Fin Tense=Prs Polarity=Pos Person=3	Aux
19	.	-	PUNCT	-	Punct

1-jadval. Iyerarxik bog'liqlik daraxt korpusining bog'liqliklar tafsiloti (1ta gap misolida)

Korpus tahlili jarayonida o'zbek tilining badiiy uslubiga xos bir qator qiziqarli sintaktik va morfologik holatlar kuzatildi. Avvalo, egasiz gaplar (subyekt tushirilgan, shaxs-son grammatik ko'rsatkichlari fe'lda mavjud) uchrashi kuzatildi. Masalan, ba'zi jumalalarda 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 jumalarda bir nechta sodda gaplar va, ham kabi bog'lovchilar

yoki -sa kabi qo'shimchalar orqali bog'lanib, murakkab gaplar hosil qilingan. Masalan: "Shu paytgacha bizda shevalar bo'yicha tadqiqotlarda ularning adabiy tildan farqi qancha ko'p ko'rsatilsa, shuncha ijobiy baholanib kelinmoqda" kabi jumalar bir vaqtning o'zida ikkita-yu uchta sodda harakatni ifodalaydi. Korpusimizda bunday hollarda ravishdoshli fe'lining bog'lanishi advcl (hol ergash gapli qo'shma gap) yoki conj (bog'lovchisiz qo'shma gap) sifatida belgilandi. Natijada, daraxt tuzilishida bir necha rootga ega bo'lmagan, balki bitta bosh fe'l atrofida birikkan struktura hosil bo'ladi (ketma-ket bog'langan fe'llar)(2-rasm). Bu o'zbek tilining polisintetik sintaksisini ko'rsatadi.



2-rasm Iyerarxik bog'liqlik daraxt korpusining vizual ko'rinishi.

Yuqoridagi kuzatishlar bizning korpusimizni rus va ingliz tillaridagi treebanklar bilan solishtirganda qiziqarli farqlarni ko'rsatadi.

Xulosa va takliflar. 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 jumalarni tahlil qilishni aynan shu korpus misollarida mashq qilishlari mumkin.

Xulosa qilib aytganda, "O'zbek tili hikoyaviy matnlari uchun iyerarxik bog'lik daraxti korpusi" – o'zbek tilini kompyuterda qayta ishlash yo'nalishida muhim qadam bo'lib, boy tilimizni formal ravishda ifodalash va ulardan amaliy yechimlar olishga xizmat qiladi.

ADABIYOTLAR

1. 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.
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. 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*.
4. 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

5. Sharipov, M., & Yuldashev, O. (2022). UzbekStemmer: Oʻzbek tili uchun qoidaviy stemming algoritmi. ALTNLP 2022 konferensiyasi materiallari, CEUR-WS, vol. 3315, sah. 137–144
6. Salaev, U. (2023). UzMorphAnalyzer: Oʻzbek tili uchun morfologik tahlil modeli (soʻzlardagi qoʻshimchalarga asoslangan). Science and Innovation, 2(1), 29–34.
7. Shuhrat Matkarim, Boljon, Toshkent, 2022-y, 318
8. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021). UZWORDNET: Oʻzbek tili uchun leksik-semantik maʼlumotlar bazasi. 11-Global WordNet konferensiyasi materiallari, sah. 8–19, Janubiy Afrika (UNISA).
9. de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies: koʻp tilli sintaktik annotatsiya tizimi. Computational Linguistics, 47(2), 255–308.
10. 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.
11. Rahmatullayev, Sh. (2006). Hozirgi adabiy oʻzbek tili (darslik). Toshkent: Universitet nashri.
12. Sharipov, M., Kuriyozov, E., Yuldashev, O., & Sobirov, O. (2023). UzbekTagger: Oʻzbek tili uchun qoidaviy soʻz turkumlari teglagichi. arXiv preprint arXiv:2301.12711.
13. Universal Dependencies Treebank for Uzbek
14. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation - ACL Anthology