UDK: 811.512.133`42

TURKIY TILLARDAGI TREEBANKLAR KLASSIFIKATSIYASI

Bozorqulova O‘g‘iloy Erkin qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
ogiloybozorqulova62@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur tadqiqotda turkiy tillar uchun yaratilgan treebanklarning tasnifi o‘rganilib, ularning lingvistik xususiyatlari hamda qo‘llanilish sohalari atroflicha tahlil qilinadi. Turkiy tillarning grammatik tuzilishi va so‘z yasash xususiyatlari va ushbu tushunchalar treebanklarga qay darajada ta’sir etishiga ham alohida to‘xtalib o‘tiladi.

Abstract. This study analyses the classification of treebanks created for Turkic languages, their linguistic features, and their scope of use. A particular focus is placed on the extent to which the grammatical structure and word formation features of Turkic languages affect treebanks.

Аннотация. В данном исследовании анализируется классификация древесных банков, созданных для тюркских языков, их лингвистические особенности и сфера использования. Особое внимание уделяется тому, в какой степени грамматическая структура и особенности словообразования тюркских языков влияют на банки деревьев.

Kalit so‘zlar: *Treebank, turkiy tillar, sintaktik tahlil, morfologik teglash, tabiiy tilni qayta ishlash, avtomatik analiz.*

Zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasida treebanklar (sintaktik korpuslar) muhim resurs sifatida ajralib turadi. Treebank – bu tabiiy tilni qayta ishlash va kompyuter lingvistikasi uchun maxsus ishlab chiqilgan korpus bo‘lib, undagi har bir gap sintaktik tahlil qilinib, daraxt strukturasida (tree structure) annotatsiya qilingan. Boshqacha qilib aytganda, treebank faqat matnlarni saqlab qolishdan tashqari, ularning grammatik tuzilishi, jumladan, so‘z turkumlari, fraza strukturalari va sintaktik bog‘lanishlari haqidagi ma’lumotlarni ham taqdim etadi. Bu resurslar zamonaviy lingvistik tahlil va NLP modellari uchun asosiy vosita sifatida qo‘llaniladi. Treebankning asosiy xususiyatlari quyidagilardan iborat: sintaktik annotatsiyalash jarayonida har bir so‘zga uning grammatik roli (masalan, ot, fe’l, sifat va boshqalar) va sintaktik bog‘lanishlar (qaysi so‘z qaysi so‘z bilan bog‘langanligi) aniqlanadi. Misol uchun, “Kitobni stolga qo‘ydim” gapini tahlil qilsak, “kitobni” - to‘ldiruvchi, “stolga” - hol,

“qo‘ydim” esa kesim sifatida belgilangan. Treebank ushbu grammatik va sintaktik ma’lumotlarni tizimli ravishda tashkil etib, lingvistik tahlillar hamda NLP tadqiqotlari uchun muhim manba vazifasini bajaradi.

Daraxt shaklidagi strukturalash jarayonida gap tarkibidagi so‘zlar o‘rtasidagi sintaktik munosabatlar daraxt shaklida ifodalanadi. Ushbu yondashuv sintaktik tahlilning aniq va tizimli tasvirini taqdim etib, har bir so‘zning grammatik roli hamda boshqa so‘zlar bilan bog‘lanish xususiyatlarini vizual ravishda aks ettiradi. Daraxt strukturalash tabiiy tilni qayta ishlash va lingvistik tahlil uchun asosiy vosita sifatida xizmat qiladi. Treebanklar constituency (bo‘laklar) yoki dependency (bog‘lanishlar) asosida qurilishi mumkin. Bunday strukturalar har bir gapning sintaktik munosabatlarini aniq va tizimli shaklda tasvirlash imkonini beradi. Treebanklar tabiiy tilni qayta ishlash sohasida juda muhim manba hisoblanib, NLP modellarini (masalan, sintaktik tahlil dasturlari va mashina tarjimasi tizimlari) o‘qitishda qo‘llaniladi. Shuningdek, ular til qonuniyatlarini o‘rganish, grammatik xatolarni aniqlash va tuzatish kabi vazifalarni amalga oshirishda keng foydalaniladi. Ushbu resurslarning amaliy ahamiyati lingvistika va kompyuter fanlari kesishmasida katta o‘rin egallaydi. Treebanklar matnlarning grammatik tuzilishini daraxt shaklida (masalan, dependency parsing yoki constituency parsing) aniqlash orqali mashinalarga tilni tushunish va generatsiya qilish imkoniyatini taqdim etadi. Shunga qaramay, dunyoda 180 milliondan ortiq so‘zlashuvchiga ega bo‘lgan turkiy tillar (jumladan, turk, o‘zbek, qozoq, uyg‘ur, qirg‘iz va boshqalar) uchun treebank resurslari nisbatan kam bo‘lib, ular tizimli klassifikatsiyaga ega emas. Ushbu tadqiqotning maqsadi turkiy tillar uchun yaratilgan treebanklarni lingvistik, texnologik va annotatsiya jihatlari asosida tasniflash, mavjud resurslarning afzalliklari va kamchiliklarini aniqlash hamda kelajakdagi loyihalar uchun tavsiyalar ishlab chiqishdan iborat.

Turkiy tillar o‘ziga xos morfologik va sintaktik xususiyatlari bilan ajralib turadi. Ushbu xususiyatlar treebank yaratishda Yevropa tillari (masalan, ingliz tili) uchun ishlab chiqilgan an’anaviy annotatsiya sxemalarini moslashtirish jarayonida qiyinchiliklar keltirib chiqaradi. Masalan, Universal Dependencies (UD) loyihasini turkiy tillarga moslashtirish bir qancha muammolarni yuzaga keltiradi, chunki bu tillarning morfosintaktik tuzilishi UD talablariga to‘liq mos kelmaydi. Natijada, turkiy tillar uchun treebanklarni alohida tasniflash va ularga moslashtirilgan annotatsiya qoidalarini ishlab chiqish zarurati paydo bo‘ladi. Shunga qaramay, turkiy tillar uchun treebanklarning yetishmasligi mintaqada NLP texnologiyalarining sekin rivojlanishiga sabab bo‘lmoqda. Misol uchun, o‘zbek tilida ishlaydigan ChatGPTga o‘xshash modellar hali cheklangan darajada faoliyat olib bormoqda, chunki ular o‘zbek tilining murakkab sintaktik tuzilishini to‘liq o‘zlashtira olmagan. Ushbu holat turkiy tillar uchun chuqur lingvistik tadqiqotlar va zamonaviy NLP resurslarini yaratish zarurligini yaqqol ko‘rsatadi. Hozirgi vaqtda

turkiy tillar ichida faqat turk tili uchun keng qamrovli treebank resurslari mavjud. Ulardan biri – Turkish Web Treebank [2], bu 1 million so‘zdan iborat korpus bo‘lib, Universal Dependencies (UD) sxemasi asosida annotatsiya qilingan. Shuningdek, METU-Sabancı Treebank [3] turk tilining morfologik tahliliga qaratilgan. Boshqa turkiy tillar, jumladan uyg‘ur va qozoq tillari uchun treebanklar faqat akademik loyihalar doirasida yaratilgan. Masalan, Kazakh Dependency Treebank [4] 30 000 so‘zdan iborat bo‘lib, asosan gazeta matnlarini o‘z ichiga oladi.

Jahonning ko‘zga ko‘ringan tillari, jumladan ingliz, nemis va fransuz tillarida treebank dasturlarini yaratish dastlab bir qancha muammolarni hal qilishni talab etgan. Turkiy tillarda treebank tizimini yaratishda esa tilning o‘ziga xos imkoniyatlari va lingvistik xususiyatlarini inobatga olgan holda ish olib borish zarur. Xususan, morfologik murakkablik va sintaktik moslikni Universal Dependencies (UD) annotatsiyasiga moslashtirish masalasi dolzarb ahamiyat kasb etadi. Shu bilan birga, turkiy tillarning morfosintaktik tuzilmalari UD talablariga to‘liq mos kelmasligi sababli bir qancha nomuvofiqliklar yuzaga keladi. Bu nomuvofiqliklar sintaktik tartib va morfologik shakllar asosida namoyon bo‘lib, treebank tizimini ishlab chiqish jarayonida maxsus yondashuvlarni talab qiladi. Chunki turkiy tillarda so‘zlar affikslar orqali ko‘plab grammatik ma’nolarni ifodalaydi. Bundan tashqari turkiy tillarda sintaktik tartib quyidagi ko‘rinishda belgilanadi: SOV – “Men [subyekt] kitob [obyekt] o‘qiyman [fe’l]. Bu kabi xususiyatlar esa treebank annotatsiyasini murakkablashtiradi. Ammo bu kabi muammolarga qaramasdan turkiy tillarga mansub ba’zi tillarda ushbu tizim yaratilgan bo‘lib, bugungi kunda tizimni rivojlantirish uchun say-harakatlar amalga oshirilmoqda. Quyida turkiy tillarda mavjud treebanklar haqida asosiy ma’lumotlar bilan to‘ldirilgan jadval keltirilgan.

1-jadval. Turkiy tillarda mavjud treebanklar tasnifi

Til	Treebank nomi	Korpus hajmi	Annotatsiya sxemasi	Mavjudligi	Xususiyatlar/ kamchiliklari
Turk tili	Turkish Web Treebank (TWT)	1 million so‘z	Universal Dependencies (UD)	Ochiq manba	Murakkab morfologiya, ko‘p qatlamli annotatsiya
Turk tili	METU – Sabanci Treebank	500 000 so‘z	LFG (Lexical-Functional Grammar)	Akademik loyiha	Morfologik tahlilga qaratilgan
Qozoq tili	Kazakh UD Treebank	30 000 so‘z	Universal Dependencies	Ochiq manba	Asosan gazeta matnlari qamrab olingan

Uyg‘ur tili	Uyghur UD Treebank	20 000 so‘z	UD + alohida morfologik belgilar	Tajribaviy	NLP vositalari to‘liq qo‘llanilmagan
Qirg‘iz tili	-	-	-	Yo‘q	Hech qanday treebanklar mavjud emas
Turkman tili	-	-	-	Yo‘q	Faqat lug‘atlar va grammatik qo‘llanmalar
Tatar tili	Tatar Corpus (Arenas et al.)	10 000 so‘z	UD (qisman)	Cheklangan	Noto‘liq annotatsiya, asosan folklor matnlari

Yuqoridagi treebanklar orasida faqat turk va qozoq tillari uchun nisbatan tizimli treebanklar mavjud. Shunga qaramay, mavjud treebanklar bir qator kamchiliklarga ega, jumladan, matn turlarining bir xilligi (asosan gazeta va adabiy matnlar bilan cheklangan) va dialektlar hamda shevalarni qamrab olmasligi. Lotin-kirill alifbosidagi tillar, xususan o‘zbek tili uchun esa Universal Dependencies (UD) annotatsiyasini moslashtirish nisbatan osonroq kechadi. UD qoidalariga ko‘ra, standart POS teglar mavjud bo‘lib, bu teglar so‘z turkumlarining nomlari orqali (masalan, noun, verb, adjective) annotatsiya qilinadi. Shuningdek, UDda so‘z tartibi SVO (subject-verb-object) formatida belgilanadi. Masalan, “I read a book” gapida I – subject (ega), read – verb (kesim), a book – object (to‘ldiruvchi) sifatida aniqlanadi. Ushbu belgilash tizimi dunyo tillarining ko‘pchiligi uchun mos keladi. Biroq, turkiy tillarning o‘ziga xos xususiyatlari bor. Turkiy tillar agglyutinativ tillar guruhiga mansub bo‘lib, bu tillar murakkab morfologik belgilarni o‘z ichiga oladi. Bundan tashqari, turkiy tillarning so‘z tartibi SOV (subject-object-verb) shaklida bo‘lib, bu dunyo tillarining ko‘pchiligidagi SVO tartibidan farqlanadi. Ushbu morfosintaktik xususiyatlar turkiy tillar uchun UD annotatsiyasini kengaytirish zaruratini ko‘rsatadi. Global tajribani turkiy tillarga tatbiq qilish orqali ham muammoni hal qilish mumkin. Masalan, boshqa agglyutinativ tillar, jumladan fin va venger tillari uchun Fin UD treebankida morfologik belgilar XPOS va UPOS annotatsiyasi orqali qo‘shilgan[1]. Japanese BCCWJ [5] korpusiga esa matn turlari (rasmiy va norasmiy) bo‘yicha segmentatsiya kiritilib, tizimdagi muammolar bartaraf etilgan. Ushbu tajribalar turkiy tillar uchun ham Universal Dependencies (UD) annotatsiyasini yaxshilashda qo‘llanilishi mumkin.

2-jadval. UD va Turkiy tillar annotatsiyalarining xususiyatlari taqqoslangan.

Universal Dependencies	**Turkiy tillar annotatsiyasi**
Standart POS teglar	Murakkab morfologik belgilar
SVO sintaktik tartibi	SOV sintaktik tartibi

Global bog‘lanishlar | Tilga xos bog‘lanishlar

Treebanklarni yaratishga qaratilgan ilmiy ishlar va maqolalarda turkiy tillar va UD annotatsiyalari orqasidagi farqlar va o‘zaro moslashtirish masalasi ilgari suriladi. Quyidagi jadvalda Universal Dependencies va Turkiy tillar annotatsiyasi o‘rtasidagi farqlar keltirib o‘tilgan.

3- jadval. Universal Dependencies va Turkiy tillar annotatsiyasi o‘rtasidagi asosiy farqlar

Xususiyat	**UD standarti**	**Turkiy tillar (Moslashtirilgan)**	**Misollar**
Morfologik annotatsiya	Oddiy Pos teglar (Noun, Verb kabi)	Qo‘shimcha morfologik belgilar (Case, Voice).	*Uyg‘ur: Noun+Case=Genitive*
Sintaktik bog‘lanishlar	Asosiy bog‘lanishlar (nsubj, obj, obl).	Tilga xos bog‘lanishlar (compound:lvc,aux:q).	*O‘zbek: compound: lvc (yordamchi fe’l).*
So‘z tartibi	SVO (Subject-Verb_Object) ga mos	SOV (Subject-Object-Verb) bilan moslashtirilgan.	*“Men kitob o‘qidim” nsubj (o‘qidim, Men).*
Affikslar	So‘z sifatida ajratilmaydi	Morfemalarga bo‘linishi mumkin.	*Turk: “okudugum” “oku+dug+um”.*
Case marking	Case xususiyati oddiy (Nom, Acc, Gen).	Batafsil case belgilari (Abl, Loc, Equ).	*Qozoq: Case = Equ*
Qo‘llanilish darajasi	100+ tilda standart	Faqat 5-6 turkiy tilda (turk, qozoq).	*Uyg‘ur UD beta versiyasida*

Yuqoridagi jadvaldan kelib chiqadiki, Universal Dependencies (UD) loyihasi global miqyosdagi tillarning turli xususiyatlarini qamrab oladi, bu esa UD tizimining odatiy morfologik belgilar to‘plamidan ham ko‘rinadi. Turkiy tillarga moslashtirilgan UD sxemasiga qo‘shimcha morfologik belgilar kiritilishi tilning o‘ziga xos xususiyatlarini hisobga olish zaruratini ifodalaydi. Masalan, **case** turkiy tillarda so‘z turkumining holatini belgilash vazifasini bajaradi, **voice** esa fe’llarning modallik belgilarini aniqlash uchun ishlatiladi. Bu kategoriyalar o‘z ichida ichki guruhlarga bo‘linishi mumkin, bu esa tilni yanada chuqurroq tahlil qilish imkonini beradi. Sintaktik tartib masalasi ham bu jarayonda muhim o‘rin egallaydi. UD standartlariga ko‘ra so‘zlar morfemalarga ajratilmaydi, biroq turkiy tillar o‘zining grammatik o‘ziga xosliklari sababli bu talabni qo‘llab-quvvatlamaydi. Agglyutinativ xususiyatga ega bo‘lgan turkiy tillarda murakkab so‘zlarni morfemalarga ajratish zarur. Masalan, qozoq tilidagi “барыпжатырмын” (borayotirman) so‘zi quyidagicha morfemalarga bo‘linadi: **бар** – fe’l (bor), - **ып** – zamon qo‘shimchasi,

- **жатыр** – hozirgi zamon qo‘shimchasi, - **мын** 1-shaxs shaxs-son qo‘shimchasi (-man). Ushbu xususiyatlar turkiy tillarning grammatik tizimini aniq va to‘liq tasvirlash uchun morfemalashni talab qiladi. Bundan tashqari, UD tizimi dunyo bo‘ylab 100 dan ortiq tillarga moslashtirilgan va faol qo‘llanilmoqda. Ammo turkiy tillar oilasidagi tillar faqat cheklangan darajada moslashtirilgan variantlaridan foydalanmoqda. Bu holat turkiy tillarning lingvistik o‘ziga xosliklarini inobatga olgan holda, UD sxemasini yanada kengaytirish zarurligini ko‘rsatadi. Bu o‘z navbatida turkiy tillar uchun mustahkam va moslashtirilgan treebank resurslarini yaratishda muhim ahamiyat kasb etadi.

Xulosa qilib aytganda, turkiy tillar uchun yaratilgan treebanklar lingvistik va kompyuter lingvistikasi nuqtayi nazaridan alohida ahamiyatga ega. Ushbu resurslar sintaktik tahlil jarayonini avtomatlashtirish, tabiiy tilni qayta ishlash tizimlarini rivojlantirish va turkiy tillarning grammatik xususiyatlarini chuqur tadqiq qilishda muhim manba bo‘lib xizmat qiladi. Turkiy tillar agglutinativ xususiyatlari bilan ajralib turgani sababli, ularga mos treebanklar so‘z shakllari, morfologik tahlil va sintaktik tuzilish nuqtayi nazaridan murakkablik kasb etadi. Mazkur tasnif turkiy tillar uchun mavjud treebanklarning tuzilishini, ularning lingvistik modelga asoslangan yondashuvlarini hamda annotatsiya tizimlarining farqlarini aniq ko‘rsatadi. Shu bilan birga, treebanklar o‘rtasidagi o‘xshashlik va farqlar tahlil qilinib, ularning tabiiy tilni qayta ishlashdagi samaradorligi baholangan. O‘rganilgan treebanklar turli standartlarga asoslangan bo‘lib, ba’zilari Universal Dependencies (UD) formatida ishlab chiqilgan, boshqalari esa milliy lingvistik an’analarga tayangan. Natijada, turkiy tillar uchun treebank yaratish jarayonida umumiy qoidalar va standartlar katta ahamiyat kasb etishi ta’kidlangan. Shu bilan birga, har bir tilning o‘ziga xos grammatik va morfologik xususiyatlarini hisobga olish zarurligi ham alohida qayd etilgan. Bunday yondashuv turkiy tillar uchun mos treebank resurslarini yaratishga imkoniyat beradi.

Foydalanilgan adabiyotlar:

1. Pyysalo S., Kanerva J., Missila A., Laipala V., Jinter F. The Finnish universal dependencies Treebank/ Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). – Finlandiya, 2015. 164-165-p.
2. Pamay T., Sulubacak U., Torunog‘lu-Selamet D., Eryigit G. The annotation process of the ITU WEB Treebank / Proceedings of LAW IX - The 9th Linguistic Annotation Workshop. – Colorado, 2015. 95-p.
3. Sulubacak U., Eryigit G., Pamay T. A revisted dependency Turkish Treebank/ Proceedings of TurCLing 2016, the 1st International Conference on Turkic Computational Linguistics. – Turkey, 2016. 4-p.

4. Tyers F., Washington J. Towards a free/open-source universal-dependency treebank for Kazakh / 3rd International Conference on Computer Processing in Turkic Languages (TURKLANG 2015). – Kazan, 2015. 3-p.

5. Omura M., Asahara M. UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese / Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). – Brussel, 2018. 117-p.