TURKIY TILLAR SINTAKSISI UCHUN KO‘P DARAJALI HAMDA IYERARXIK KOMPYUTERLI MODELLAR

Norov Abdisait Muradovich

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti doktoranti (DSc), Toshkent, O‘zbekiston.
E-mail: abdusaidnorov@gmail.com
ORCID: 0000-0002-0458-2023

K E Y W O R D S

turkiy tillar, sintaksis, kompyuterli modellashtirish, NLP, Dependency Tree, Constituency Tree, iyerarxik model, agglyutinativ tillar

A B S T R A C T

Maqolada turkiy tillar sintaktik tuzilishini kompyuterli modellashtirishning nazariy va amaliy asoslari tahlil qilingan. Agglyutinativ tillar xususiyatlaridan kelib chiqqan holda sintaktik tahlil jarayoni uchun ko‘p darajali hamda iyerarxik model taklif etilgan. Modelda morfema, so‘z, so‘z birikmasi, gap va matn darajalari o‘zaro bog‘langan holda ifodalanadi. Shuningdek, “Dependency Grammar” va “Constituency Grammar” yondashuvlarining turkiy tillar uchun qo‘llanilishi tahlil qilinib, o‘zbek tili misolida sintaktik daraxtlar qurish algoritmi ishlab chiqilgan.

KIRISH.

So‘nggi yillarda tabiiy tillarni qayta ishlash sohasida neyron tarmoqlar va katta til modellarining rivojlanishi milliy tillarning sintaktik modellarini takomillashtirishga bo‘lgan ehtiyojni oshirdi. Turkiy tillar agglyutinativ xususiyatga ega bo‘lganligi sababli ularning sintaktik tuzilishini formal modellashtirish analitik tillarga nisbatan murakkabroq hisoblanadi. O‘zbek, qozoq, qirg‘iz, turk va uyg‘ur tillarida grammatik munosabatlarning asosiy qismi qo‘shimchalar orqali ifodalanadi. Bu esa sintaktik tahlil jarayonida morfologik va sintaktik darajalarning chambarchas bog‘liqligini ta’minlaydi. Shu nuqtai nazardan turkiy tillar uchun ko‘p darajali hamda iyerarxik kompyuterli modellarni ishlab chiqish muhim ilmiy vazifalardan biri hisoblanadi.

Turkiy tillar dunyo tillari orasida agglyutinativ tuzilishi bilan ajralib turadi. Bu tillarda grammatik ma’no asosan qo‘shimchalar orqali ifodalanadi va har bir qo‘shimcha muayyan grammatik vazifani bajaradi. Shu sababli turkiy tillarning sintaktik qurilishi analitik tillarga nisbatan o‘ziga xos xususiyatlarga ega.

Kompyuter lingvistikasi nuqtai nazaridan sintaksis tilning morfologik va semantik sathlari o‘rtasidagi bog‘lovchi bo‘g‘in hisoblanadi. Sintaktik tahlil jarayonida so‘zlar o‘rtasidagi grammatik munosabatlar, ularning gap tarkibidagi vazifalari va iyerarxik bog‘lanishlari aniqlanadi.

Zamonaviy NLP tizimlarida sintaktik tahlil quyidagi vazifalarni hal qilishda muhim ahamiyatga ega:

- mashinaviy tarjima;
- avtomatik referatlash;
- axborot-qidiruv tizimlari;
- savol-javob tizimlari;
- katta til modellari;
- intellektual ta’lim platformalari va h.k.

Turkiy tillar uchun mazkur vazifalarni amalga oshirishda sintaktik modellashtirish asosiy bosqichlardan biri hisoblanadi.

ADABIYOTLAR TAHLILI.

Sintaktik modellashtirish sohasida bir qator olimlar tomonidan muhim tadqiqotlar amalga oshirilgan. Jumladan, Noam Chomsky [1], [2] tomonidan generativ grammatika nazariyasi ishlab chiqilgan bo‘lsa, Lucien Tesnière [3] dependency grammar konsepsiyasini taklif qilgan. Joakim Nivre [4], [5] esa Universal Dependencies loyihasi doirasida zamonaviy sintaktik tahlil usullarini rivojlantirgan.

(cc) BY

Turkiy tillar uchun sintaktik tahlil bo‘yicha tadqiqotlar turk, qozoq va o‘zbek tillari korpuslari asosida amalga oshirilgan bo‘lsa-da, ko‘p darajali iyerarxik modellar masalasi hali yetarlicha o‘rganilmagan.

TADQIQOT METODOLOGIYASI.

Kompyuterli sintaktik tahlilning nazariy asoslari formal grammatikalar nazariyasiga borib taqaladi. Formal grammatika tilning sintaktik qurilishini matematik va algoritmik vositalar orqali ifodalash imkonini beradi. Noam Chomsky tomonidan taklif etilgan generativ grammatika nazariyasiga ko‘ra, har qanday gap cheklangan sonli qoidalar orqali hosil qilinishi mumkin. Bu yondashuv keyinchalik kompyuterli sintaktik tahlil tizimlarining nazariy asosiga aylangan.

Formal grammatika odatda quyidagi to‘rtta komponentdan tashkil topadi: noterminal simvollar to‘plami (N); terminal simvollar to‘plami (T); qayta yozish qoidalari (P) va boshlang‘ich simvol (S). Buni G=(N,T,P,S) kabi ifoda etsak, ushbu model turkiy tillar sintaksisini formallashtirish uchun ham muhim nazariy asos bo‘lib xizmat qiladi.

So‘nggi yillarda tabiiy tillarni avtomatik qayta ishlashda Dependency Grammar yondashuvi keng qo‘llanilmoqda. Lucien Tesnière [3] nazariyasiga ko‘ra, gap tarkibidagi har bir so‘z boshqa bir so‘zga tobe holatda bo‘ladi. Natijada daraxtsimon struktura hosil bo‘ladi. Masalan, “inson mehnatda toblanadi” gapida “toblanadi” – markaziy tugun, “inson” – ega va “mehnatda” – to‘ldiruvchi bo‘lib xizmat qilmoqda.

Turkiy tillarda so‘z tartibi nisbatan erkin bo‘lganligi sababli dependency modellari constituency modellariga nisbatan samaraliroq hisoblanadi. Ayniqsa, Universal Dependencies loyihasi doirasida o‘zbek tili uchun ham sintaktik bog‘lanishlarni standartlashtirish ishlari olib borilmoqda.

Constituency Grammar yondashuvida gap alohida so‘zlar emas, balki sintaktik birliklar (constituents) yig‘indisi sifatida qaraladi. Masalan, “yozuvchi yangi kitobni yozdi” jumlasi quyidagi tarkibiy qismlarga ajratiladi:

- NP (Noun Phrase) – “yozuvchi”;
- NP (Noun Phrase) – “yangi kitobni”;
- VP (Verb Phrase) – “yozdi”.

Bu yondashuv grammatik strukturalarni chuqur tahlil qilish imkonini bersa-da, agglyutinativ tillarda morfologik ma’lumotlarni alohida hisobga olishni talab qiladi.

Tabiiy tillarni qayta ishlash sohasida zamonaviy yondashuvlar tilni bir nechta o‘zaro bog‘langan sathlar tizimi sifatida qarashni taqozo etadi. Turkiy tillar uchun bunday sathlar quyidagilardan iborat: morfema sathi, so‘z sathi, so‘z birikmasi sathi, gap sathi va matn sathi. Har bir yuqori sath o‘zidan pastki sath elementlari asosida shakllanadi. Masalan:

[morfema] → [so‘z] → [so‘z birikmasi] → [gap] → [matn].

Bunday yondashuv tilning tabiiy iyerarxik tuzilishini aks ettiradi va kompyuterli modellashtirish uchun qulay arxitekturani ta’minlaydi.

Mavjud sintaktik tahlil tizimlarining aksariyati ingliz tili kabi analitik tillar uchun yaratilgan. Turkiy tillarning morfologik boyligi va agglyutinativ xususiyati esa ularni to‘g‘ridan-to‘g‘ri qo‘llash imkonini cheklaydi. Shu sababli turkiy tillarda “morfologik ma’lumot”, “sintaktik munosabatlar”, “semantik kontekst” va “iyerarxik bog‘lanish” kabilar bir vaqtning o‘zida hisobga olinadigan maxsus modellarga ehtiyoj seziladi.

Yuqorida ko‘rib chiqilgan nazariy yondashuvlar va turkiy tillarning lingvistik xususiyatlaridan kelib chiqqan holda mazkur tadqiqotda turkiy tillar sintaksisi uchun quyidagi ko‘p darajali (besh pog‘onali) hamda iyerarxik kompyuterli model taklif etiladi.

1-daraja. Morfemalar qatlami. Bu yerda har bir morfema mustaqil tugun sifatida qaraladi, masalan: kitob + lar + im + ni.

2-daraja. So‘zlar qatlami. Bunda morfemalar birlashib so‘z tugunlarini hosil qiladi, masalan: [kitoblarimni] [o‘qidim].

3-daraja. So‘z birikmalari qatlami. Bunda sintaktik munosabatlar shakllanadi, masalan: (kitoblarimni o‘qidim).

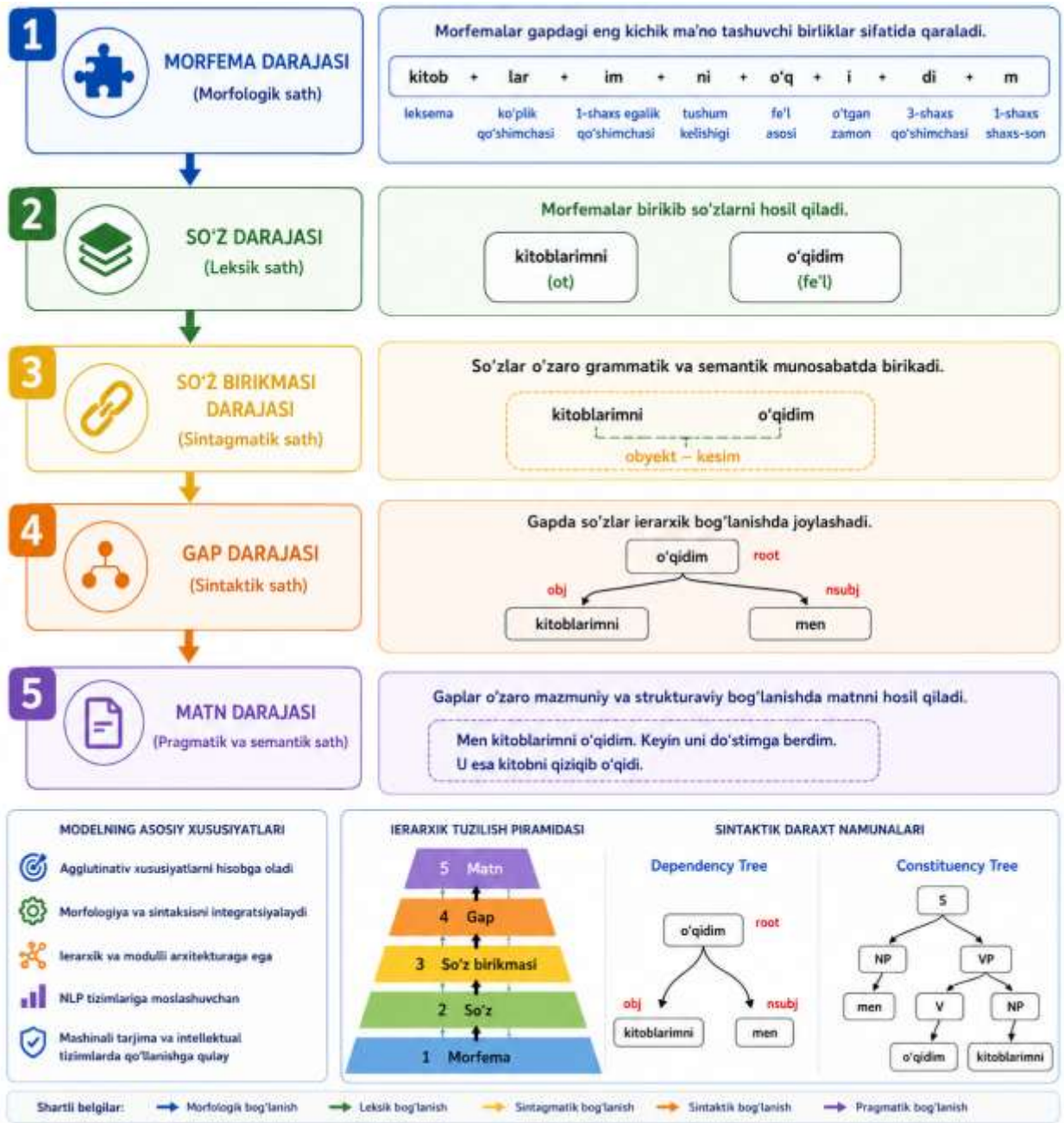
4-daraja. Gap qatlami. Bu bosqichda Dependency Tree quriladi, masalan: “men kitoblarimni o‘qidim” jumlasi uchun T=(V,E) kabi

(cc) BY

ifodani yozish mumkin, bu yerda: V – sintaktik tugunlar to‘plami; E – sintaktik munosabatlar to‘plami.

5-daraja. Matn qatlami. Bunda bir nechta gaplar o‘rtasidagi semantik va sintaktik bog‘lanishlar ifodalanadi.

Yuqorida ko‘rganimiz singari, turkiy tillardagi istalgan gap uchun iyerarxik sintaktik daraxt modelini osongina qurish mumkin (1-rasmga qarang):

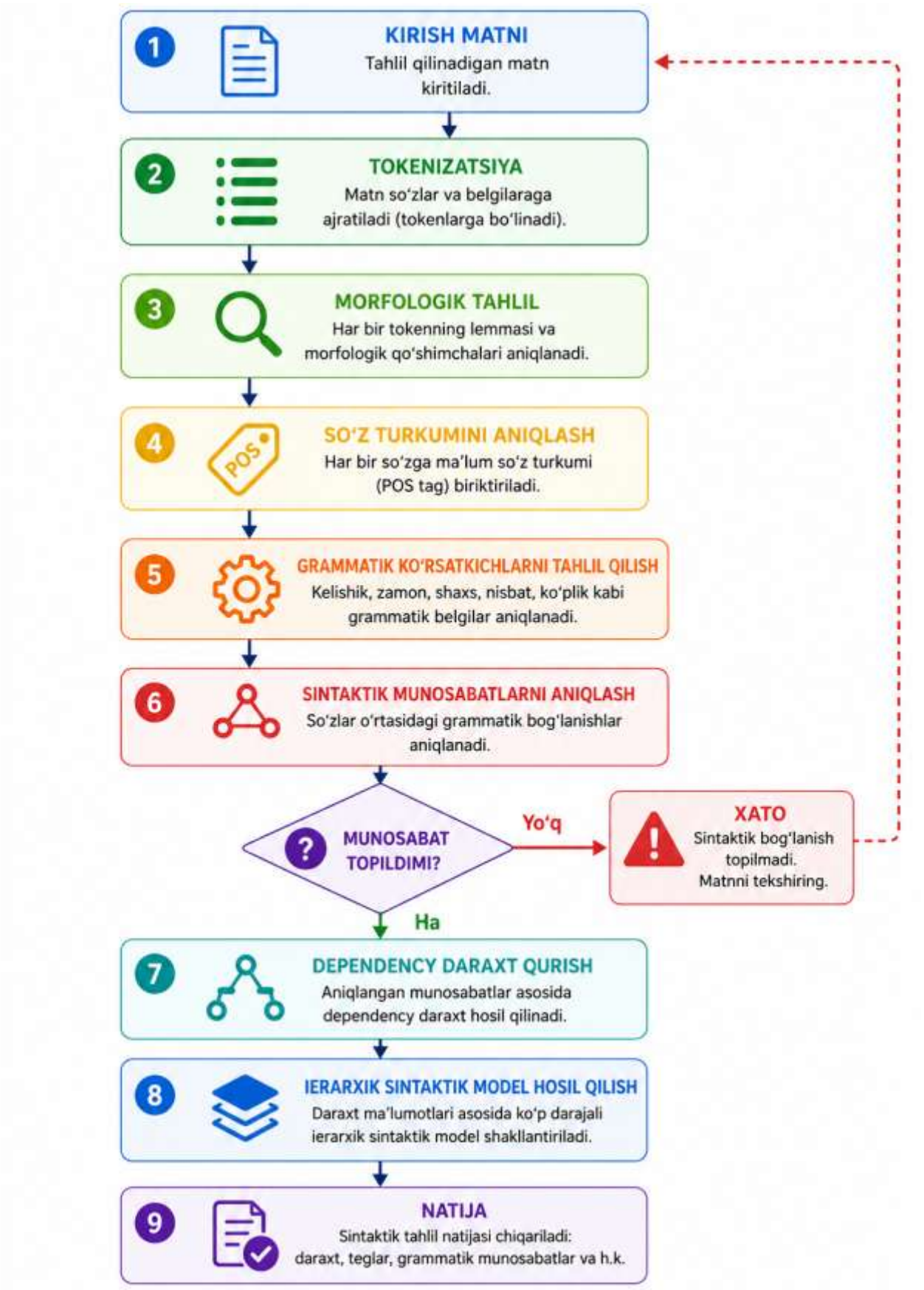


1-rasm. Turkiy tillar sintaksisi uchun ko‘p darajali hamda iyerarxik model

Agar iyerarxik sintaktik daraxt modeliga e’tibor qaratadigan bo‘lsak, ushbu daraxtda kesim gapdagi markaziy tugun hisoblanadi, ega va to‘ldiruvchi esa bevosita kesimga bog‘lanadi, shuningdek, agar gapda hol ham ishtirok etayotgan bo‘lsa, u ham ikkinchi darajali bog‘lanishni tashkil etadi. Haqiqatan ham qat’iy qoidaga amal qilinadigan bo‘lsa, ega guruhi va ikkinchi darajali bo‘laklar har doim kesimga ergashib keladi.

Shunday qilib, turkiy tillar sintaksisiga oid taklif etilayotgan ko‘p darajali hamda iyerarxik kompyuterli model uchun algoritmning umumiy sxemasi quyidagicha bo‘ladi (2-rasmga qarang).

(cc) BY



2-rasm. Turkiy tillar sintaksisini ko'p darajali iyerarxik model asosida kompyuterli tahlil qilish algoritmi.

(cc) BY

MUHOKAMA VA NATIJALAR.

Tadqiqot doirasida turkiy tillarning agglyutinativ xususiyatlari, sintaktik munosabatlarning o‘ziga xosligi hamda zamonaviy tabiiy tillarni qayta ishlash tizimlarining talablari tahlil qilindi. Tahlillar shuni ko‘rsatdiki, mavjud sintaktik modellarning aksariyati analitik tillar uchun ishlab chiqilgan bo‘lib, ular turkiy tillarda uchraydigan morfologik murakkabliklarni to‘liq qamrab ololmaydi. Taklif etilgan ko‘p darajali va iyerarxik model mazkur muammoni hal qilishga qaratilgan bo‘lib, unda til birliklari morfema, so‘z, so‘z birikmasi, gap va matn darajalari bo‘yicha yagona arxitektura asosida tashkil etildi. Bunday yondashuv morfologik va sintaktik ma’lumotlarni integratsiyalash imkonini berdi hamda turli darajalar o‘rtasidagi bog‘lanishlarni formal ko‘rinishda ifodalashga xizmat qildi.

Ishlab chiqilgan algoritmning asosiy afzalligi shundaki, u sintaktik tahlil jarayonini bosqichma-bosqich amalga oshiradi. Dastlab tokenizatsiya va morfologik tahlil orqali til birliklari aniqlanadi, keyin esa ularning grammatik xususiyatlari hisobga olingan holda sintaktik munosabatlar shakllantiriladi. Natijada dependency daraxt quriladi va ushbu daraxt asosida ko‘p darajali iyerarxik model hosil qilinadi.

Taklif etilgan modelning nazariy tahlili uning quyidagi ustunliklarini ko‘rsatdi:

- turkiy tillarning agglyutinativ xususiyatlarini hisobga olishi;
- morfologiya va sintaksisning o‘zaro integratsiyalashgan holda ishlashi;
- dependency va constituency yondashuvlarining afzalliklarini birlashtirishi;
- sintaktik tahlil natijalarining iyerarxik ko‘rinishda saqlanishi;
- NLP tizimlarida qayta foydalanish imkoniyatining yuqoriligi.

Shuningdek, taklif etilgan arxitektura kelgusida o‘zbek tili uchun sintaktik korpuslar, avtomatik parserlar va katta til modellarini yaratishda nazariy asos sifatida xizmat qilishi mumkin.

Tadqiqot natijalari shuni ko‘rsatadiki, ko‘p darajali model turkiy tillardagi sintaktik munosabatlarni formallashtirishda an’anaviy bir sathli modellarga nisbatan samaraliroq hisoblanadi. Ayniqsa, morfologik ma’lumotlarning sintaktik tahlil jarayoniga jalb qilinishi tahlil aniqligini oshirishga xizmat qiladi.

XULOSA VA TAKLIFLAR.

Ushbu tadqiqotda turkiy tillar sintaksisini kompyuterli modellashtirishning nazariy asoslari tahlil qilinib, ko‘p darajali hamda iyerarxik arxitekturaga ega bo‘lgan yangi model taklif etildi. Model morfema, so‘z, so‘z birikmasi, gap va matn darajalarini yagona tizimda birlashtirib, ular o‘rtasidagi sintaktik munosabatlarni formal ifodalash imkonini yaratadi. Tadqiqot natijalari asosida quyidagi xulosalarga kelindi:

- turkiy tillarning agglyutinativ xususiyati sintaktik tahlil jarayonida morfologik ma’lumotlarni hisobga olishni talab qiladi.
- Dependency Grammar turkiy tillardagi sintaktik munosabatlarni modellashtirish uchun samarali vosita hisoblanadi.
- ko‘p darajali iyerarxik yondashuv til birliklari o‘rtasidagi bog‘lanishlarni to‘liqroq aks ettiradi.
- taklif etilgan algoritm sintaktik tahlil jarayonini formallashtirish va avtomatlashtirish imkonini beradi.
- model o‘zbek tili va boshqa turkiy tillar uchun intellektual NLP tizimlarini yaratishda qo‘llanilishi mumkin.

Olib borilgan tadqiqotlar hamda erishilgan natijalar asosida kelgusidagi tadqiqotlar uchun quyidagi yo‘nalishlar taklif etiladi:

- o‘zbek tilining dependency treebank korpusini kengaytirish;
- taklif etilgan model asosida avtomatik sintaktik parser yaratish;
- neyron tarmoqlar va transformer arxitekturalari bilan integratsiyalash;
- turkiy tillar uchun Universal Dependencies standartiga mos ma’lumotlar bazasini shakllantirish;
- katta til modellari (LLM) uchun milliy lingvistik resurslar yaratish.

(cc) BY

Shunday qilib, taklif etilgan ko‘p darajali va iyerarxik model turkiy tillar sintaksisini kompyuterli tahlil qilishning yangi konseptual yondashuvini shakllantirib, tabiiy tillarni qayta ishlash tizimlari, mashinaviy tarjima, avtomatik tahlil va intellektual axborot tizimlarini yaratish, milliy til texnologiyalarini rivojlantirish uchun muhim nazariy va amaliy ahamiyat kasb etadi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI:

1. Chomsky N. Syntactic Structures. – The Hague: Mouton & Co., 1957. – 117 p.
2. Chomsky N. Aspects of the Theory of Syntax. – Cambridge, MA: MIT Press, 1965. – 251 p.
3. Tesnière L. Éléments de syntaxe structurale. – Paris: Klincksieck, 1959. – 670 p.
4. Nivre J. Dependency Grammar and Dependency Parsing // MSI Report. – Växjö University, Sweden, 2005. – 32 p.
5. Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D. Universal Dependencies v1: A Multilingual Treebank Collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). – Portorož, Slovenia, 2016. – P. 1659–1666.
6. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – 3rd ed. – Stanford University Draft, 2025. – 690 p.
7. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge, MA: MIT Press, 1999. – 680 p.
8. Kübler S., McDonald R., Nivre J. Dependency Parsing. – San Rafael, CA: Morgan & Claypool Publishers, 2009. – 105 p.
9. Erdal M. A Grammar of Old Turkic. – Leiden: Brill, 2004. – 575 p.
10. Johanson L., Csató É. Á. The Turkic Languages. – London: Routledge, 1998. – 608 p.

(cc) BY