

SINTAKTIK PARSING TURLARI

Elov Botir Boltayevich,

Texnika fanlari falsafa doktori, dotsent

elov@navoiy-uni.uz

ToshDO‘TAU

Abdullayeva Oqila Xolmo‘minovna,

doktorant DSc

abdullayevaoqila@gmail.com

ToshDO‘TAU

Annotatsiya. NLPning muhim yo‘nalishlaridan biri matnlarni sintaktik tahlil qilishdir. So‘nggi tadqiqotlar natijasida tillarda sintaktik tahlilni amalga oshiruvchi parser dasturlar tuzildi. Gap darajasidagi sintaktik ma‘lumotlarni ifodalash uchun ko‘plab murakkab qoidalarni tahlil qilish kerak. Ushbu maqolada sintaktik tahlil turlari ichida keng tarqalgan dependency (tobelik munosabati tahlili) va constituency (tarkibiy tuzilma tahlili) parsing turlari muhokama qilindi. Kontekstdan mustaqil grammatika (CFG), ya‘ni constituency tahlil (yoki birikma tuzilmasi asosidagi tahlil), gapning sintaktik tuzilmasidagi iyerarxik so‘z birikmasini treebank orqali tasvirlashga asoslanadi va bu yondashuv ilk tadqiqot obyekti bo‘lib kelmoqda. Bog‘liqlik (yoki dependency) grammatikasi esa sintaktik va semantik tahlil uchun keng qo‘llaniladigan yana bir mashhur model bo‘lib, unda so‘zlar bevosita bog‘liqlik munosabatlari orqali birikadi va ushbu bog‘liqliklarga ular orasidagi sintaktik yoki semantik munosabatlarni ko‘rsatuvchi teglar belgilanadi. Har ikkala parsing turi tillarda gaplarni sintaktik tahliliga tatbiq qilinib, korpuslarda qo‘llanilmoqda, lekin resurs va grammatik tuzilishiga ko‘ra olinadigan natijalar farqliligi kuzatilmoqda. Maqolada parsing turlarining grammatik xususiyatlari o‘zaro taqqoslandi.

Abstract. One of the important areas of NLP is the syntactic analysis of texts. As a result of recent research, parser programs that perform syntactic analysis in languages have been created. To express syntactic information at the sentence level, it is necessary to analyze many complex rules. This article discusses the most common types of syntactic analysis: dependency (dependence relation analysis) and constituency (structural analysis) parsing. Context free grammar (CFG), that is, constituency analysis (or analysis based on the structure of a phrase), is based on describing a hierarchical phrase in the syntactic structure of a sentence through a tree bank, and this approach is the first object of research. The grammar of dependence (or dependency) is another popular model widely used for syntactic and semantic analysis, in which words are connected through direct dependency relations, and tags are assigned to these connections, indicating the syntactic or semantic relations between them. Both types of parsing are applied to the syntactic analysis of sentences in languages and are used in corpora, but there is a difference in the results

obtained according to the resource and grammatical structure. In the article, the grammatical features of the types of parsing were compared.

Аннотация. Одним из важных направлений НЛП является синтаксический анализ текстов. В результате последних исследований были созданы программы-параметры, выполняющие синтаксический анализ в языках. Для выражения синтаксической информации на уровне предложения необходимо проанализировать многие сложные правила. В данной статье обсуждаются наиболее распространенные типы синтаксического анализа: анализ зависимости (анализ зависимости отношений) и синтаксический анализ (структурный анализ). Контекст-свободная грамматика (KFG), то есть анализ составляющих (или анализ, основанный на структуре словосочетания), основан на описании иерархического словосочетания в синтаксической структуре предложения через банк деревьев, и этот подход является первым объектом исследования. Грамматика зависимости (или зависимости) - это еще одна популярная модель, широко используемая для синтаксического и семантического анализа, в которой слова связаны посредством прямых зависимых отношений, и этим связям присваиваются теги, указывающие на синтаксические или семантические отношения между ними. Оба типа синтаксического анализа применяются к синтаксическому анализу предложений в языках и используются в корпусах, но есть разница в результатах, полученных в соответствии с ресурсом и грамматической структурой. В статье сравнивались грамматические особенности типов синтаксиса.

Kalit so‘zlar: *sintaktik parsing, parser, dependency parsing, constituency parsing, teg.*

NLPda parsing – matnning grammatik tuzilishini, gap bo‘laklarining vazifasini va matn tarkibidagi so‘zlar o‘rtasidagi munosabatlarni aniqlash uchun tahlil qilish jarayoni hisoblanadi. Tilshunoslik nuqtayi nazaridan qaralganda, tahlilning asosiy vazifasi so‘zlarning qanday birikib gap hosil qilishi va gap tuzilishini belgilovchi qoidalarni tahlil qilish bo‘lsa, boshqa tomondan, NLP amaliyotlari nuqtayi nazaridan tahlil mashina tarjimai, savol-javob tizimlari, axborot ajratib olish, sentiment tahlil va matn yaratish kabi ko‘plab vazifalar uchun muhim hisoblanadi.

Har bir tabiiy tilning asosiy tuzilishi mavjud bo‘lib, odatda bu “grammatika” deb ataladi. Grammatika o‘z navbatida morfologiya va sintaksis kabi bo‘limlarga ajratilib o‘rganiladi. Sintaksisning asosiy g‘oyasi shundan iboratki, so‘zlar birlashib, so‘zlar guruhi yoki so‘z birikmasi kabi tarkibiy qismlarni hosil qiladi va ular bir

butun birlik sifatida harakat qiladi. Ushbu tarkibiy qismlar o‘zaro birlashib, yirikroq birliklarni va oxir-oqibatda gaplarni hosil qilishi mumkin. Tahlil jarayoni ko‘pincha maxsus grammatik qoidalarga tayangan holda amalga oshiriladi, bu grammatik qonunlar sintaktik tuzilmalarni aniq shakllantirishga xizmat qiladi.

Sintaktik parsing turlari bo‘yicha amalga oshirilgan tadqiqot natijalarini tahlil qilganimizda, dependency parsing, constituency parsing, kombinator kategoriyali grammatika (CCG), bosh so‘zga asoslangan birikma tuzilmasi grammatikasi (HPSG), leksik-funksional grammatika (LFG), mavhum ma’no tasviri (AMR), minimal rekursiv semantika (MRS), universal tushunchaviy-kognitiv annotatsiya (UCCA) hamda mantiqiy tahlilga yo‘naltirilgan bir nechta tahlil turlarini o‘rganish mumkin [1]. Ushbu grammatik yo‘nalishlarning barchasi so‘nggi yillarda tadqiq qilinib, ulardan ayrimlari kuchli neyron tarmoqlar va kontekstual oldindan o‘rgatilgan so‘z tasvirlarining (representations) paydo bo‘lishi sababli tez sur‘atlar bilan rivojlanmoqda. Mazkur tadqiqotimizda faqatgina dependency va constituency parsing masalasiga to‘xtalamiz.

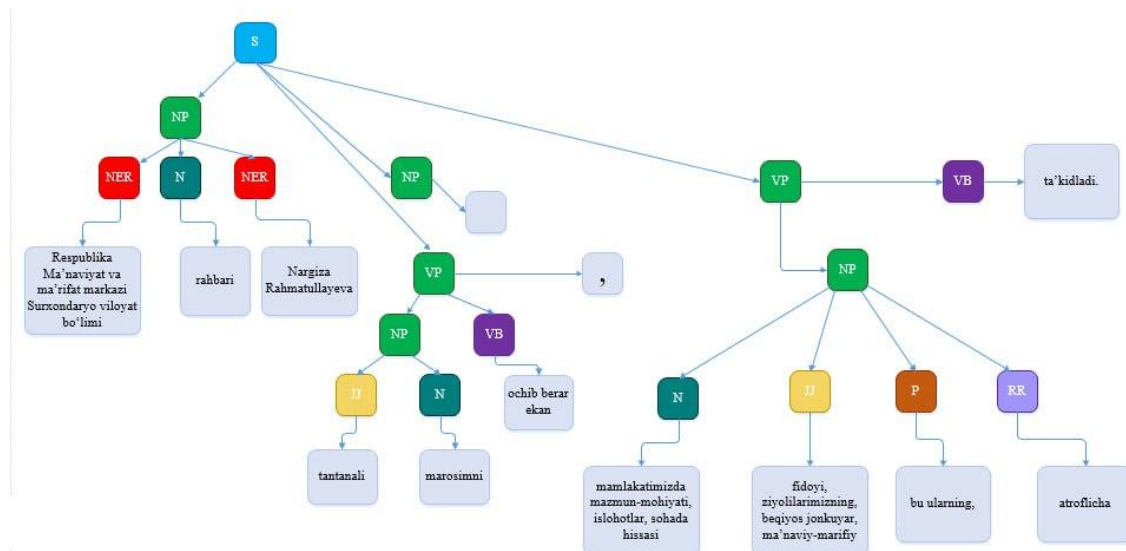
Dependency parsing - bu tilshunoslik va tabiiy tilni qayta ishlash sohasidagi muhim tadqiqot mavzusidir, chunki u so‘zlar orasidagi grammatik bog‘lanishlarni aniqlashga yordam beradi. Bu yondashuvning rivoji, tilshunoslikni chuqurroq tushunishga imkon beradi va tabiiy tilni qayta ishlashda (NLP) samarali foydalaniladi. Ilgari ilmiy ishlar qaramlikni yoki bog‘liqlikni aniq tahlil qilishda grammatika qoidalariga asoslangan metodlarga tayangan bo‘lsa, hozirgi kunda statistik va sun‘iy intellekt asosidagi usullar va modellar ko‘plab tillarda yuqori samarali natijalar ko‘rsatmoqda. Ayniqsa, transformer arxitekturalari (masalan, BERT va GPT) tilning sintaktik va semantik strukturalarini tushunishda katta yordam beradi.

Konstituent tahlil (constituency parsing) sintaktik tahlilning asosiy vazifalaridan biri hisoblanadi. 1-rasmda konstituent daraxtining namunasi ko‘rsatilgan bo‘lib, bu yerda tugunlar konstituent segmentlari, ya’ni frazalar sifatida namoyon bo‘lgan. Konstituent tahlilning asosiy maqsadi ushbu frazalarni belgilash va ularning o‘zaro munosabatlarini aniqlashdan iboratdir. Bu parsing turi bo‘yicha amalga oshirilgan tadqiqotlarni tahlil qilganimizda, gap tuzilishida qat’iy so‘z tartibi mavjud tillarda ko‘proq mavjudligini kuzatish mumkin. Erkin so‘z tartibli tuzilmalarga ega tillarda frazalarni ajratish muammolidir. Chunki turkiy tillarda erkin so‘z tartibi mavjud (SOV asosiy bo‘lsa-da), bu esa fraza tuzilmasini qat’iy daraxtlarda ifodalashni qiyinlashtiradi. Dependency tahlil ko‘proq mos keladi: chunki u har bir so‘z orasidagi bog‘lanishni ko‘rsatadi va agglyutinatив tillar uchun ancha samaraliroq hisoblanadi. Shu bois, Universal Dependencies (UD) loyihasi ham turkiy tillar uchun aynan dependency parsing asosida korpuslar yaratmoqda. Quyidagi gapni konstituent parse daraxtida tahlil qilamiz: *Respublika Ma’naviyat va ma’rifat markazi Surxondaryo viloyat bo‘limi rahbari Nargiza Rahmatullayeva*

tantanali marosimni ochib berar ekan, mamlakatimizda ma'naviy-ma'rifiy sohada amalga oshirilayotgan islohotlar, ularning mazmun-mohiyati haqida atroflicha to'xtalib, bu borada jonkuyar, fidoyi ziyolilarimizning hissasi beqiyos ekanini ta'kidladi.

O'zbek tili morfoanalizatorida (<https://uznatcorpara.uz/>) POS teglangan:

Respublika Ma'naviyat va ma'rifat markazi[[NER]] *Surxondaryo viloyat bo'limi*[[NER]] *rahbari*[[N]] *Nargiza Rahmatullayeva*[[NER]] *tantanali*[[JJ]] *marosimni*[[N]] *ochib berar*[[VB]] *ekan*[[VB]], *mamlakatimizda*[[N]] *ma'naviy-ma'rifiy*[[JJ]] *sohada*[[N]] *amalga oshirilayotgan*[[IB]] *islohotlar*[[N]], *ularning*[[P]] *mazmun-mohiyati*[[N]] *haqida*[[II]] *atroflicha*[[RR]] *to'xtalib*[[VB]], *bu*[[P]] *borada*[[II]] *jonkuyar*[[JJ]], *fidoyi*[[JJ]] *ziyolilarimizning*[[JJ]] *hissasi*[[N]] *beqiyos*[[JJ]] *ekanini*[[VB]] *ta'kidladi*[[VB]].



1-rasm. Constituent daraxti namunasi

Parse daraxti misolida ko'rish mumkinki, gap tuzilmasini tarkibiy qismlarga ajratib tahlil qilish murakkablik tug'diradi.

Bu sintaktik parsing turlari yondashuviga ko'ra turli grammatik munosabatlarni aniqlashga xizmat qiladi [4].

Constituency grammar (tarkibiy tuzilma grammatikasi) - Constituency grammatika, shuningdek, birikma strukturasi grammatikasi deb ham ataladi. Buning sababi, u tarkibiy birikma munosabatiga asoslanadi. Bu bog'lanish grammatikasining qarama-qarshi turidir.

Constituency grammatikada tarkiblar har qanday so'z, so'zlar guruhi yoki birikmalar bo'lishi mumkin. Constituency grammatikadagi asosiy maqsad – har qanday gapni uning tarkiblariga ajratish va ularning xususiyatlari asosida tuzishdir.

Tarkibiy tuzilma grammatikasi va tarkib munosabatining xususiyatlari:

– barcha tegishli ramkalar gap tuzilishini tarkibiy birikma munosabati orqali ko‘radi;

– tarkibiy munosabatni hosil qilish uchun biz lotin va yunon grammatikalaridagi subyekt-predikat bo‘linishini qo‘llaymiz;

– tarkibiy tuzilma grammatikasida gap tuzilishini ot iborasi (NP) va fe‘l iborasi (VP) nuqtayi nazaridan o‘rganamiz. Constituency grammatika har qanday gapni uchta tarkibga ajratishi mumkin: subyekt, predikat va obyekt.

– Tarkibiy tuzilma grammatikasi kontekstsiz va dependency grammatikalarining cheklovlarini hal qilishda yaxshiroq ishlaydi:

– bu yondashuv usuli tilga xos emas, bu esa bir xil modelni bir nechta tillarda ishlatishni yoki tillar o‘rtasida almashishni osonlashtiradi, shu bilan boshqa ikki grammatik turdagi ko‘p tillilik muammosini hal qiladi;

– bu tahlil turida tahlil daraxtini ishlatganligi sababli gapning tarkiblari o‘rtasidagi iyerarxik munosabatni ifodalashni osonlashtiradi va bu boshqa tasvirlash grammatikalariga nisbatan ko‘proq tushunarli va intuitivdir;

– tarkibiy tuzilma grammatikasi boshqa usullarga nisbatan soddaroq va amalga oshirish osonroq bo‘lib, masalan, bog‘lanish grammatikasi kabi formalizmga qaraganda tadqiqotchilar va amaliyotchilar uchun qulayroqdir;

– bu tahlil usuli xatolarga chidamli va shovqinli yoki to‘liq bo‘lmagan ma’lumotlarni boshqarishda yaxshi ishlaydi, shuningdek, so‘z birikmalari yoki jumalarni bog‘lashda qo‘llaniladigan birlashtiruvchi vosita sifatida koordinatsiyani boshqarishda yaxshiroq ishlaydi.

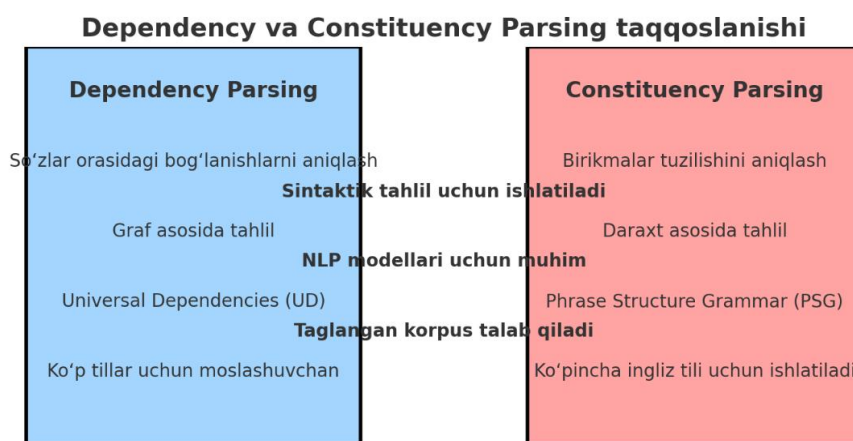
Dependency (Bog‘liqlik) grammatikasi. Bog‘liqlik grammatikasi Constituency grammatikaning qarama-qarshi turidir va bog‘lanish munosabatiga asoslanadi. U Constituency grammatikaga qarama-qarshi, chunki u birikma tugunlarini o‘z ichiga olmaydi. Bog‘lanish grammatikasi gapdagi so‘zlar bir-biriga bog‘lanish munosabatini tahlil qiladi. Bu so‘zlar bog‘lanish grammatikasida yo‘naltirilgan havolalar orqali bir-biriga ulanadi. Kesim gap tuzilmasining markazi sifatida qaraladi. Bog‘lanish grammatikasi gapdagi so‘zlarni ularning bog‘lanishlariga qarab tartiblaydi. Har bir sintaktik birlik kesimga nisbatan yo‘naltirilgan belgilash orqali bog‘lanadi. Bu sintaktik birliklar **bog‘lanishlar** deb ataladi. Gapdagi so‘zlardan biri ildiz sifatida ishlaydi va boshqa barcha so‘zlar, o‘sha so‘zni hisobga olmaganida, to‘g‘ridan-to‘g‘ri yoki bilvosita ildiz bilan o‘z bog‘lanishlari orqali bog‘lanadi. Ushbu bog‘lanishlar gapdagi so‘zlar o‘rtasidagi munosabatlarni ifodalaydi va dependency grammatikasi so‘zlar o‘rtasidagi struktura va semantik bog‘lanishlarni aniqlash uchun ishlatiladi.

Bog‘lanish grammatikasining ba’zi cheklovlari mavjud, ularni quyida tushunib chiqamiz:

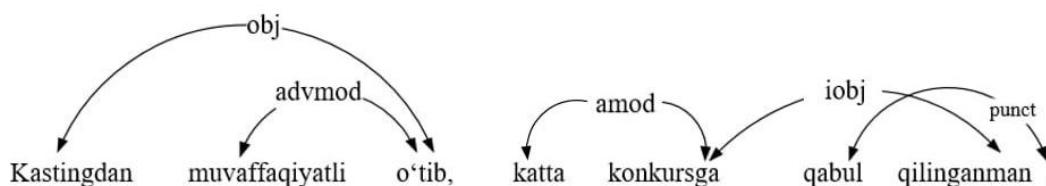
- ma’lumotlarni annotatsiya qilish: Bog‘lanish tahlilini o‘rganish modelini o‘qitish uchun belgilangan ma’lumotlar kerak bo‘ladi, bu esa vaqtni talab qiladi va olish qiyin;
- uzoq masofadagi bog‘lanishlarni boshqarish: Ba’zi hollarda, gapdagi so‘zlar o‘rtasidagi bog‘lanishlar juda uzoqda bo‘lishi mumkin, bu esa gapning grammatik tuzilishini aniq olishni qiyinlashtiradi;
- ellipsis va koordinatsiyani boshqarish: Bog‘lanish grammatikasi, shuningdek, so‘zlar o‘rtasidagi to‘g‘ridan-to‘g‘ri bog‘lanishlar bilan tushuntirilmagan hodisalarni, masalan, ellipsis va koordinatsiyani boshqarishda qiyinchiliklarga duch keladi, bu hodisalar odatda Constituency grammatika orqali ko‘rib chiqiladi.

Dependency grammatikasining cheklovlarini yengish uchun Constituency grammatika qo‘llanilishi mumkin, bu esa kamroq kuchli bo‘lsa-da, lekin ko‘proq tushunarli va amalga oshirish osonroqdir. Shuningdek, Constituency va Dependency grammatikalari birgalikda qo‘llanadigan gibrid yondashuvdan foydalanish foydali bo‘lishi mumkin.

Tarkibiy tuzilma grammatikasi va bog‘liqlik grammatikasini taqqoslash: Constituency grammatika so‘zlarni birikma va jummalarga guruhlashga e’tibor qaratadi, dependency grammatikasi esa har bir so‘z o‘rtasidagi munosabatlarga qaratiladi. Har bir so‘z gapda bog‘lanish grafida tugun sifatida ifodalanadi va tugunlar o‘rtasidagi qirralar so‘zlar o‘rtasidagi grammatik munosabatlarni ifodalaydi. Bog‘lanish grammatikasi ayrim tillar va NLP vazifalari uchun ko‘proq kuchli bo‘lishi mumkin, chunki u so‘zlar o‘rtasidagi munosabatlarni aniqroq tushuntiradi, ammo buni amalga oshirishda yanada murakkab va kamroq intuitivdir. Constituency grammatika esa ko‘proq tushunarli va amalga oshirish osonroq, lekin ifodalashda kamroq kuchli bo‘lishi mumkin. Quyidagi jadvalda ikki parsing turining o‘xshash va farqli jihatlari aniq ko‘rsatilgan.

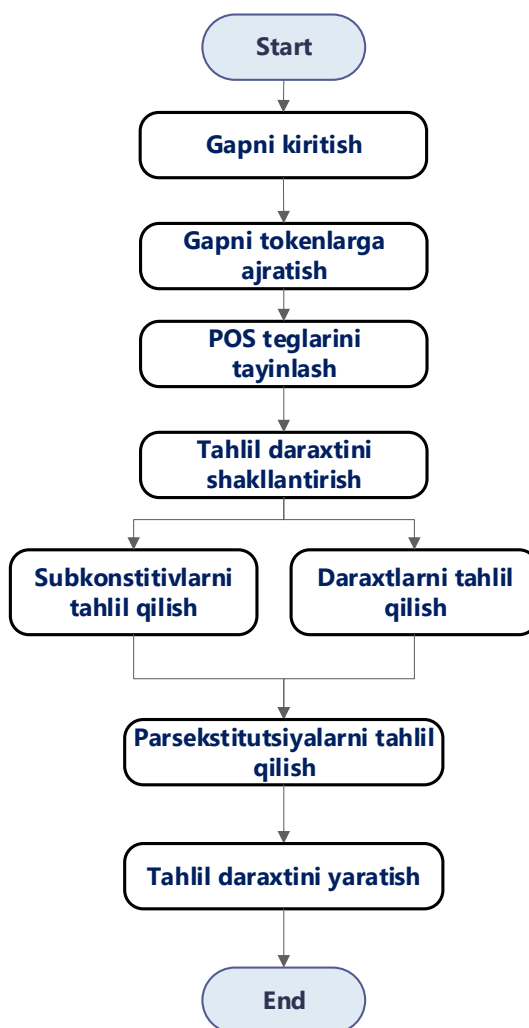


Bog‘liqlik tahlili sintaktik va semantik tahlil uchun ishlab chiqilgan bo‘lib, u bi-leksikallashgan bog‘liqlik grammatikasiga, ya’ni ikki so‘z munosabatiga asoslanadi. Bu yerda barcha sintaktik va semantik hodisalar bi-leksikallashgan bog‘liqliklar orqali ifodalanadi. 2-rasmda bog‘liqlik tahlilining namunaviy daraxti keltirilgan. Tahlilda UD universal dependency teglaridan foydalanilgan. Turli bog‘liqlik analizatorlarini baholashda asosiy mezon sifatida bog‘liqlik aniqligi qo‘llaniladi, bu esa teglanmagan birikma ko‘rsatkichi (UAS – *Unlabeled Attachment Score*) va teglangan birikma ko‘rsatkichi (LAS – *Labeled Attachment Score*) bilan o‘lchanadi. Boshlang‘ich bosqichlarda bog‘liqlik tahlili daraxt shaklida, ya’ni proyektiv yoki no-proyektiv tuzilmalar doirasida cheklangan edi. So‘nggi yillarda esa bog‘liqlik grafi asosidagi tahlilga qaratilgan bir qancha tadqiqotlar olib borilmoqda. Bir tomondan, dastlabki bog‘liqlik daraxtlari asosan sintaktik yo‘naltirilgan bo‘lsa, bugungi kunda so‘zlar orasidagi semantik munosabatlarga e’tibor kuchaymoqda.



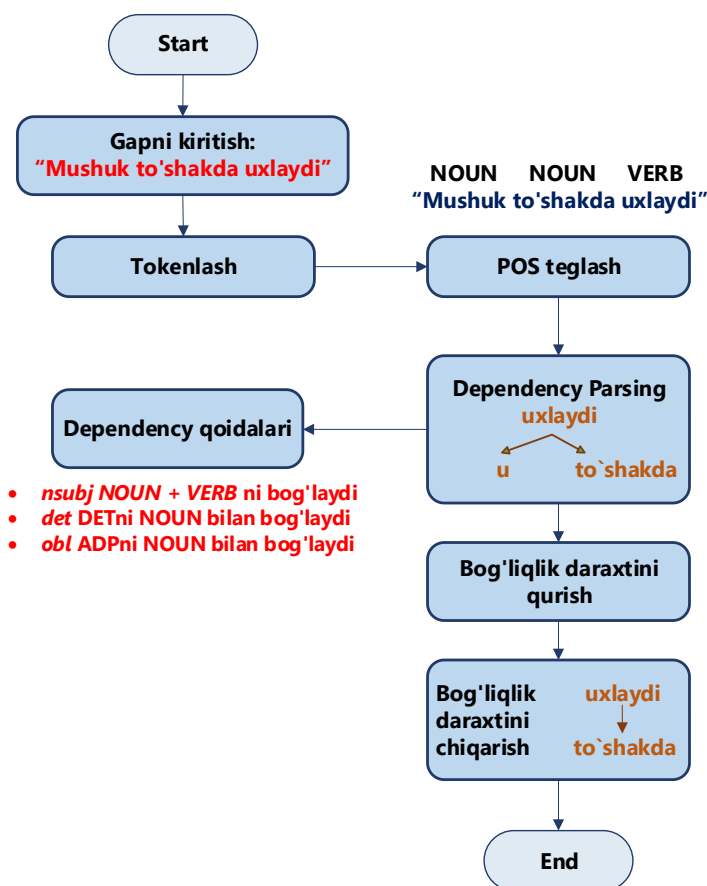
2-rasm. Dependency daraxti namunasi

Tarkibiy tuzilma tahlili Chomsky iyerarxiyasiga binoan qat’iy qoidalarga tayangani bois, qat’iy so‘z tartibli tillarda (masalan, ingliz tilida SVO) yaxshi ishlaydi, chunki grammatik munosabatlar asosan so‘zlarning tartibi bilan ifodalanadi. Ammo erkin so‘z tartibli tillarda bir fikrni turlicha tartibda ifodalash mumkin bo‘lganligi tufayli, an’anaviy tuzilmaviy grammatikalar murakkablashadi. Bog‘liqlik va tarkibiy tuzilma tahlil usullarida gaplarni tahlil qilish jarayoni ham farq qiladi.



3-rasm. Constituency parsing asosida sintaktik tahlil qilish jarayoni

Amaliy jihatdan, bugungi kunda ko‘pchilik zamonaviy parserlar bog‘liqlik paradigmasidan foydalanadi yoki dependency daraxtlarni hosil qiladi. Bunga sabab – Universal Dependencies kabi standartlarning joriy qilinishi va bog‘liqlikning kross-lingvistik (cross-lingual generalization) moslashuvchanligidir. Bu holat esa, ayniqsa, tabiiy tilni qayta ishlash (NLP) sohasida muhim bo‘lib, ko‘p tillilik (multilingualism) va kam resursli tillarni qo‘llab-quvvatlashda katta rol o‘ynaydi. O‘zbek tili erkin so‘z tartibiga ega tillardan biri bo‘lganligi uchun, oxirgi tadqiqotlarda bog‘liqlik tahlili qo‘llash samarali ekanligi xususida yozilmoqda. Chunki constituent daraxtini qurish uchun o‘zbek tilida qat’iy qoidalar asosida birikmalarni aniqlash mushkuldir. Chunki gap bo‘laklari doim bir o‘rinda kelishi yoki kelmasligi ham mumkin. Shuningdek, gaplar egali yoki egasiz bo‘lishi mumkin. Shu sababli ham bog‘liqlik grammatikasi o‘zbek tilini tahlil qilishda qulay usul bo‘lmoqda.



4-rasm. Dependency parsing asosida sintaktik tahlil qilish jarayoni

Xulosa. Grammatika – bu to‘g‘ri tuzilgan jummalarni hosil qilish uchun qoidalar sifatida aniqlanadi va tabiiy tillarda muloqotda ishlatiladigan sintaktik qoidalarni belgilashda muhim rol o‘ynaydi. Sintaksis grammatikani tashkil etuvchi tarmoq bo‘lib, tilning jumlar tuzilishini va shakllanishini o‘rganadi, ya’ni so‘zlar qanday qilib birikmalar, bandlar va jummalarni hosil qilish uchun tartiblanishini belgilaydi.

Context free grammar – bu tilni tanib olish uchun ishlatiladigan formal grammatika bo‘lib, unda ishlab chiqarish qoidalari yordamida terminal va non-terminal belgilarni almashtirish orqali tilning satrlarini aniqlash ko‘rsatiladi. Constituency grammatika - bu tilning jumlar tuzilishini tahlil qilishda ishlatiladi va jumlar ma’lum grammatik funksiyaga ega bo‘lgan kichik birliklarga - tarkiblarga ajratilishi mumkinligi g‘oyasiga asoslanadi. Bog‘lanish (Dependency) grammatikasi - bu NLPda so‘zlar o‘rtasidagi bog‘lanishlar yordamida jumlaning grammatik munosabatlarini tahlil qilish uchun ishlatiladigan usul bo‘lib, so‘zlar o‘rtasidagi bog‘lanishlar asosida jumla tuzilishini shakllantiriladi. O‘zbek tilida CFG (kontekstsiz grammatika) va bog‘liqlik grammatikalari asosida, kross-domen va kross-lingvistik muhitlarda parsing masalalarini yechish imkoniyati mavjud. Bundan tashqari, qo‘shma modellar (joint models) va parser qo‘llanmalari (parser

applications) nuqtayi nazaridan, ko'p vazifali o'rganish (multi-task learning) va oldindan o'rgatilgan (pretraining) arxitekturalar asosida ham o'zbek tili matnlarini sintaktik tahlil qiluvchi dasturlar ishlab chiqilishi va samarali natijalarga erishish mumkin.

Foydalanilgan adabiyotlar:

1. Meishan Zhang. A Survey of Syntactic-Semantic Parsing Based on Constituent and Dependency Structures. Science China Press and Springer -Verlag Berlin Heidelberg, 2017.
2. Zong Chengqing. Statistical Natural Language Processing. - Tsinghua University Press, 2013.
3. Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd (draft) ed.). Prentice Hall PTR, 2019.
4. <https://www.scaler.com/topics/nlp/introduction-to-grammar-in-nlp/>
5. James Henderson. Discriminative training of a neural network statistical parser. In Proceedings of ACL, pages 95–102, 2004.
6. Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, and et al. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In Proceedings of CoNLL 2009: Shared Task, pages 1–18, 2009.
7. Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Hajic, and Zdenka Uresova. Semeval 2015 task 18: Broadcoverage semantic dependency parsing. In Proceedings of SemEval, pages 915–926, 2015.