

IYERARXIK TAHLIL KORPUSINING ASOSIY TUSHUNCHALARI

Raximboyeva Xulkar G'ayratovna

hulkar.rahimboeva@gmail.com

Annotatsiya. Ushbu maqolada kompyuter lingvistikasi doirasida iyerarxik bog'liqlik korpusining asosiy tushunchalari, o'zbek tili uchun jahon standartiga mos yo'riqnoma ishlab chiqish va gapda tobelik, bosh asosiy negiz qismlar qanday ajratilishi ko'rsatib o'tiladi.

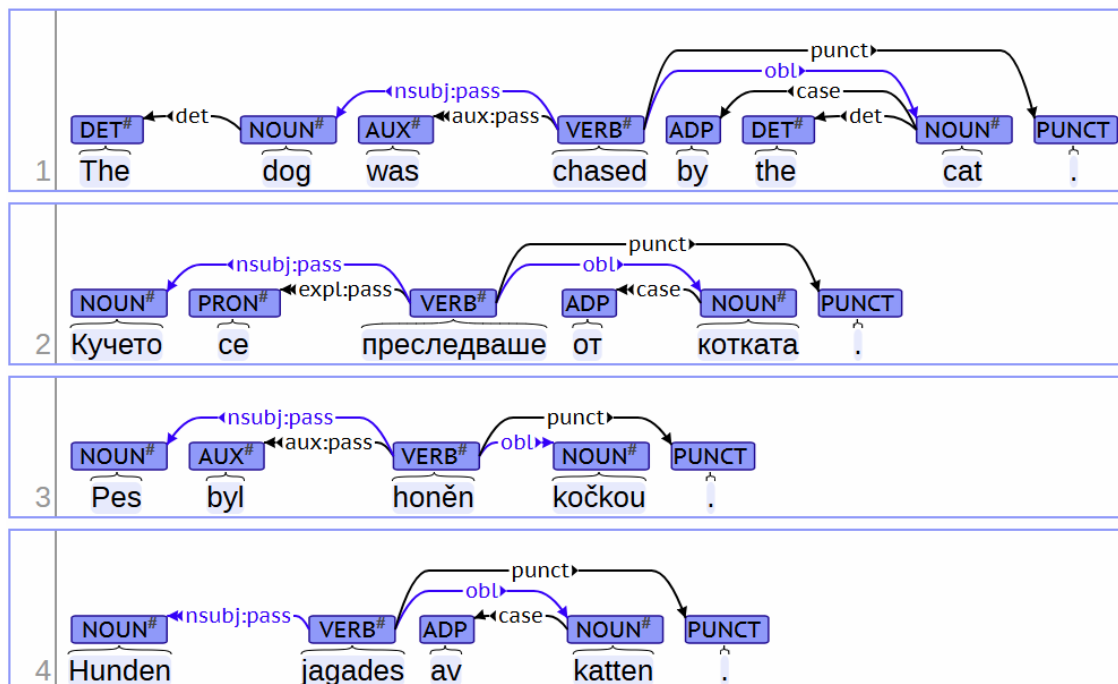
Kalit so'zlar: iyerarxiya, tobelik, bog'lanish, tabiiy tillar jarayoni, modellashtirish, negiz.

O'zbek tilining xalqaro ilmiy va texnologik maydonda rivojlanishi uchun uni tabiiy tilni qayta ishlash (NLP/Natural Language Processing) resurslar/korpuslari orqali kengaytirish muhimdir. Hozirda iyerarxik bog'liqlik daraxtini qurish uchun sintaktik va morfologik jihatdan tahlil qilingan korpuslarning yetishmovchiligi mavjud bo'lib, bu tilni sun'iy intellekt tizimlariga integratsiya qilishni qiyinlashtiradi. Iyerarxik bog'liqlikning sintaktik tahlil so'zlarning grammatik tuzilishini aniqlash orqali qidiruv tizimlari, mashina tarjimasini, nutqni qayta ishlash va inson tilini chuqur anglash kabi tabiiy tilni qayta ishlash texnologiyalarini rivojlantirishda muhim ahamiyat kasb etadi.

O'zbek tili chuqur agglutinatив xususiyatlari sababli kompyuter lingvistikasi uchun murakkab hisoblanadi. Shuning uchun o'zbek tiliga moslashtirilgan iyerarxik bog'langan sintaktik korpus yaratish lingvistik tadqiqotlar, til o'rgatuvchi dasturlar va sun'iy intellekt ilovalarining rivojlanishini jadallashtiradi. Bu tadqiqot nafaqat tilshunoslik balki texnologik rivojlanish nuqtayi nazaridan ham dolzarb va zarur bo'lib, o'zbek tilini jahonning rivojlangan tillari qatoriga qo'shilishiga va raqamli dunyoga olib chiqish imkonini beradi.

Iyerarxik bog'liqlik daraxti korpusi yoki Dependency Treebank – tabiiy tildagi gaplarning sintaktik tuzilishini aks ettiruvchi belgilangan ma'lumotlar to'plami bo'lib, unda so'zlar o'zaro bog'lanish munosabatlariga ega daraxtsimon tuzilishda taqdim

etiladi. Bunday korpus til birliklari o'rtasidagi sintaktik bog'liqliklarni aniqlash, jumlar tarkibini formal modellashtirish va tabiiy tilni qayta ishlash (NLP) algoritmlarini rivojlantirish uchun asosiy manba hisoblanadi. Iyerarxik bog'lanishlar har bir so'z uchun bosh (head) va bog'liq (dependent) elementlar orqali belgilanib, ularning sintaktik rollari ko'rsatiladi,



1.rasm. Turli tillarda iyerarxik bog'lanishlar.

Rasmda, inglizcha, bolgarcha, chexcha va shvedcha tillardagi misollarda ko'rsatilgan korpusni O'zbek tili uchun qurish maqsadimiz bo'lib. Ushbu misollarda passiv fe'l, atoqli otlar(nominal ot) va majhul nisbat o'z ichiga olgan asosiy grammatik munosabatlar bir xil bo'lsa-da, aniq grammatik ifodalanishi turlicha bo'ladi.

Bundan xulosa qilishimiz mumkinki iyerarxik bog'liqlik daraxti korpusi har bir tilning xususiyatlarini inobatga olgan holda alohida ishlab chiqilishi zarur.

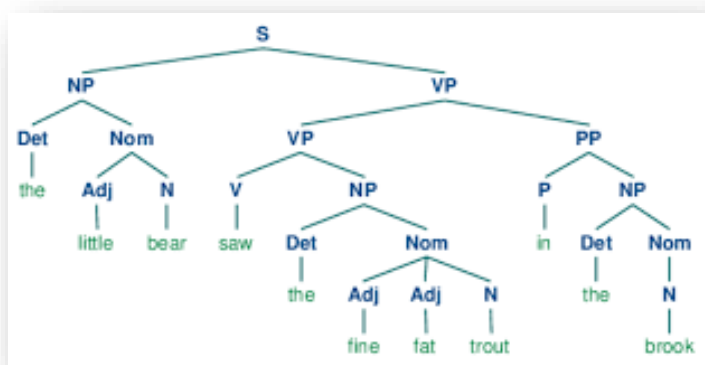
1. Dependency Treebank (Iyerarxik bog'liqlik daraxti korpusi) – gap tuzilishini sintaktik bog'liqliklar asosida aks ettiruvchi, maxsus annotatsiya bilan boyitilgan lingvistik korpus hisoblanadi. Ushbu korpus lingvistik tadqiqotlar va tabiiy tilni qayta ishlash (NLP) vazifalarida juda keng qo'llaniladi

2. Bog'liqlik (Dependency) – sintaktik jihatdan ikki so'z o'rtasida mavjud bo'lgan grammatik munosabat. Bu munosabatlar odatda bosh (head) va unga bog'liq (dependent) so'zlar orasida aniqlanadi.

3. Head (bosh) – sintaktik jihatdan asosiy ma'no va grammatik xususiyatlarni o'zida jamlagan, gapning tuzilishini boshqaruvchi so'z.

4. Tobelik(dependent) – bosh so'zga bog'langan va undan grammatik ma'no hamda funksiyani oluvchi so'z.

1. Iyerarxik bog'liqlik korpusini yaratishda annotatsiyalar odatda Universal Dependencies¹ (UD) nomli global loyiha tomonidan ishlab chiqilgan universal qo'llanmalar asosida amalga oshiriladi. Bunda e'tiborlisi shundaki, butun jahon tan olgan standard universal yo'riqnoma qoidalariga bo'ysungan holda va har bir tilning xususiyatlarini inobatga olgan holda alohida yo'riqnoma ishlab chiqilishi kerak. Quyida o'zbek tili uchun jahon standartiga mos yo'riqnoma ishlab chiqildi: Bu orqali tilshunoslar va tabiiy tilni qayta ishlovchi sun'iy intellekt injeneri mutaxassislari tillararo solishtirma tadqiqotlarni osonlik bilan olib borish imkoniyatini yaratdi. Tobelik daraxti (shajara) modeliga eng yaqin model bevosita ishtirokchilar (immediate constituents) sintaktik modeli hisoblanadi. Bevosita ishtirokchilar metodiga ko'ra, tayanch nuqta konstruksiya hisoblanadi. Konstruksiya tarkibidan ishtirokchilar va bevosita ishtirokchilar ajratiladi. Konstruksiya atamasi ostida ma'noli qismlarning ketma-ket munosabatidan tashkil topgan butunlik tushuniladi. Ishtirokchilar konstruksiya tarkibiga kirgan so'z yoki birikmalardir. Muayyan konstruksiyaning bevosita shakllanishida ishtirok etgan bir yoki bir nechta ishtirokchilarga bevosita ishtirokchilar deyiladi. Masalan, A 'lochi talabalar o'z vazifalarini aniq va puxta bajaradilar.²



¹ <https://universaldependencies.org>

² Xolmanova Z.T., Kompyuter lingvistikasi.-Toshkent:Asian Book House. 2020. 65-b.

Universal bogʻliqlik daraxti korpusi va universal yoʻriqnoma:

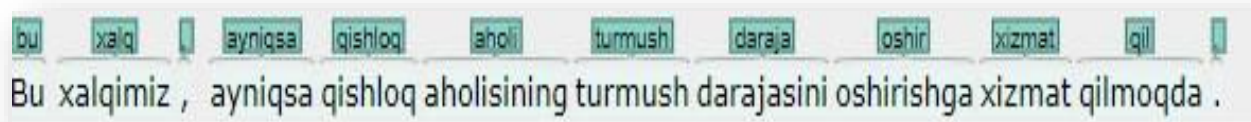
Ushbu universal standartlarning asosiy xususiyatlari quyidagilardan iborat:

*POS-teglash (soʻz turkumlarini belgilashning umumiy tizimi).

*Gapdagi soʻzlarni negizlash(lemmatization)

*Morfologik annotatsiyalar uchun universal standart, bu orqali barcha tillardagi morfologik xususiyatlarni bir xil tizimda aks ettirish mumkin boʻladi.

*sintaktik bogʻliqlik teglarini bogʻlab chiqish (nsubj, obj, root, va boshqa koʻplab bogʻliqlik turlari).



Universal iyerarxik daraxti korpusining xalqaro tajrida yaratilishi:

- Butun dunyo tajribasida iyerarxik bogʻliqlik korpusi odatda Universal Dependencies loyihasining standartlariga mos ravishda yaratiladi.

- Annotatsiya jarayoni bir nechta bosqichlardan iborat boʻlib, bu jarayon quyidagicha amalga oshiriladi:

Birinchi navbatda kerakli matnlar tanlab olinadi va ular gaplarga ajratiladi (segmentatsiya).

- Har bir soʻz uchun POS-teglash va morfologik xususiyatlarini annotatsiya qilish amalga oshiriladi.³

- Keyingi bosqichda sintaktik annotatsiya jarayoni bajarilib, soʻzlar orasidagi bogʻliqliklar belgilanadi.

³ Raxmanov X.E., Sunatov J.T., OʻZBEK TILI KOMPYUTER LINGVISTIKASI YOʻNALISHIDA OLIB BORILGAN ILMIY TADQIQOTLAR.-Toshkent: “KOMPYUTER LINGVISTIKASI: MUAMMOLAR, YECHIM, ISTIQBOL, Xalqaro ilmiy-amaliy konferensiya, Vol. 1 №. 01 (2022). 2022.224-224-bb.
www.tadqiqotlar.uz

- Oxirgi bosqich sifatida annotatsiya qilingan ma'lumotlar sifatini tekshirish va xatolarni aniqlash ishlari amalga oshiriladi. Bunda sun'iy intellekt modellaridan foydalaniladi.
- Yaratilgan korpuslar ko'pincha CoNLL-U (.conllu), XML va JSON kabi kompyuter formatlarda saqlanadi va tarqatiladi.

FOYDALANILGAN ADABIYOTLAR:

1. <https://universaldependencies.org>
2. Xolmanova Z.T., Kompyuter lingvistikasi.-Toshkent:Asian Book House. 2020. 65-b.
3. Raxmanov X.E., Sunatov J.T., O'ZBEK TILI KOMPYUTER LINGVISTIKASI YO'NALISHIDA OLIB BORILGAN ILMIY TADQIQOTLAR.-Toshkent: "KOMPYUTER LINGVISTIKASI: MUAMMOLAR, YECHIM, ISTIQBOL, Xalqaro ilmiy-amaliy konferensiya, Vol. 1 №. 01 (2022). 2022.224-224-bb.