DOI: https://doi.org/10.5281/zenodo.14041967

TIL KORPUSIDA IFODALANGAN LINGVISTIK AXBOROT DASTURLARIDAN FOYDALANISH

Xidirov Otabek
f.f.f.d (PhD),dotsent. Jizzax davlat pedagogika universiteti

Annotatsiya: *Ushbu maqolada korpuslar razmetkasida ifodalangan lingvistik axborotlardan foydalanish hamda matnga avtomatik ishlov berish, sintaktik razmetkalashning dasturiy ta'minoti tahlil ostiga olingan.*

Kalit so'zlar: *Morfologik axborot, sintaktik axborot, korpus, dasturiy ta'minot, matn razmetkasi, sintaktik tahlil dasturlari, korpus razmetkasi.*

Annotatsiya: *V state analiziruetsya ispolzovanie lingvisticheskoy informatsii, v'rajennoy v korpuse i programmnom obespechenii dlya avtomaticheskoy obrabotki teksta, sintaksicheskogo razresheniya.*

Klyuchev'e slova: *Morfologicheskaya informatsiya, sintaksicheskaya informatsiya, korpus, programmnoe obespechenie, razmetka teksta, programm' sintaksicheskogo analiza, razmetka tela.*

Annatation: *This article analyzes the use of linguistic information expressed in the case of case, as well as the software for automatic processing of text syntactic notation.*

Key words: *Morhological information, syntactic information, corpus, software, text markup, syntactic analysis programs, corpus marking.*

Gapning sintaktik belgisi shajara daraxti bilan ifodalanadi. Bu erda har bir o‘q “hokim”dan “tobe”ga yo‘naltiriladi hamda sintaktik munosabatlardan birining nomi bilan belgilanadi[1]. An’anaviy holda tarkibda sintaktik guruh hamda tarkibiy qismlar mavjud emas; aslida, shajara daraxtining har qanday “tupi” guruh deb hisoblanishi mumkin, uning tepasi “buta”ning tashqi aloqalarida uning vakili sifatida ishlaydi. Odatda, daraxt tuzilishidagi tugunlar soni jumladagi so‘zlar soniga teng. Bir tomondan, istisno holatini keltirib chiqaruvchi hodisalar ham mavjud: ayrim so‘z shakllari zanjiri leksik birlikni bildirganda, yuqoridagi qoidaga amal qilinmaydi, istisno holati yuzaga keladi. Bunday paytda, so‘z soniga qaraganda strukturada a’zo soni kamayadi. Boshqa tomondan, biror bir so‘z strukturaga “yopishtirilishi lozim” bo‘lsa yoki real matnda mavjud bo‘lmagan so‘z strukturaga kiritilishi kerak bo‘lsa ham, matndagi so‘z shakllari zanjir bilan bir xil bo‘lmasligi mumkin. (Otam vrach bo‘lib ishlardi, onam esa o‘qituvchi //Otam vrach bo‘lib ishlardi, onam esa o‘qituvchi BO‘LIB ISHLARDI). Ta’kidlash kerakki, teglash jarayonida har qanday qayta ishlangan matnning leksik, sintaktik omonimiyasi to‘liq echimini topadi. Agar buni avtomatik amalga oshirish imkoni bo‘lmasa (tizim tomonidan tuzilgan, izohlovchi yoki muharrirning fikriga ko‘ra, jumlada uchraydigan tuzilma mos kelmasa), albatta qo‘l bilan tuzatiladi.

Teglovchi mutaxassis ham omonimiyani hal qila olmaydigan alohida holatlarda (masalan, *qasddan qichqiriq bo‘lsa*, *ruh kabi ajoyib impuls mavjud*), jumla bir nechta tuzatish bilan ta’minlanishi mumkin. Razmetkalangan matnlar korpusi biror-bir tilning alohida lug‘ati bilan uzviy bog‘liq emasligi sababli muallif so‘zning aniq leksik ma’nosini ma’lum tarzda, noaniqlik va leksik omonimiya bilan belgilash g‘oyasidan voz kechadi. Masalan, *mexanik*1 (shaxs oti) va *mexanik*2 (fizik holat) so‘zi, *qalb*1 (yurak) hamda *qalb*2 (noto‘g‘ri, egri) hech qanday indeks bilan ta’minlanmagan. Istisnolar ba’zi “ishchi” so‘zlar(xususan, old qo‘shimchalar)ga oid bo‘lib, ularning leksik ma’nosi korpus hujjatlarida tasvirlangan.

[1]Богуславский И.М., Григорьев Н.В., Григорьева С.А., Иомдин Л.Л., Крейдлин Л.Г., Фрид Н.Е. Разработка синтаксически размеченного корпуса русского языка // http://cl.iitp.ru/bibitems/corpus_SPB.pdf

Demak, matnga avtomatik ishlov berish annotatsiyalangan korpusdan foydalanganda aniqroq, xatosiz amalga oshirilishi oydinlashadi. Korpusning sintaktik teg(izoh)langan qismini sinab ko‘rishni boshlagan birinchi turdagi dastur – ETAP-3 tizimida rus tilidan ingliz tiliga tarjima qilinganida sintaktik noaniqlikni avtomatik hal qilish aniq natija bergan. Bundan kelib chiqadiki, matnni avtomatik qayta ishlash uchun lingvistik ta’minot va dasturiy tizim ishlab chiqilishi talab etiladi.

Sintaktik tahlil algoritmi ishlab chiqilganda qo‘shimcha filtr yaratish ham talab etilgan: 2-4 a’zodan tashkil topgan ushbu vosita tahlil qilinayotgan gapni potensial tarmoqlar vositasida tahlildan o‘tkazadi. Bunday tajriba natijasini korpusning keyingi qismini qurishda ham qo‘llash mumkin, chunki yangi, avtomatik ravishda qurilgan gaplarni tahlil qilish yanada osonlashadi.

O.I.Babina, N.Yu.Dyuminlarning ta’kidlashicha, har bir korpus razmetkasi asosida til nazariyasi yotadi, korpus asosidagi har qanday xulosa shu konsepsiyaga asoslangan holda chiqariladi[1]. Har bir tadqiqotchi tilni modellashtirishda ob’ektiv/sub’ektiv sabablarga ko‘ra ma’lum nuqtai nazarni qo‘llab-quvvatlashi mumkin: masalan, til strukturasidan kelib chiqib, tilni formallashtirish nazariyasi ma’lum bir tilga nisbatan qo‘llanadi, boshqa tilga to‘g‘ri kelmasligi mumkin; til modeli, ko‘pincha, hatto bir til doirasida ham, undagi xilma-xillik va ko‘pma’nolilikni aks ettirolmaydi. Tadqiqot maqsadidan kelib chiqqan holda, uncha katta bo‘lmagan tadqiqiy korpuslarni lingvistik annotatsiyalash (razmetkalash) chuqur sintaktik va semantik razmetkani qamrab olishi, shuningdek, faqatgina morfologik izoh (razmetka komponenti) bilan cheklanib qolishi ham mumkin. Katta korpuslar razmetkasida ifodalangan mufassal lingvistik axborot nihoyatda katta mehnatni talab qiladi. Tadqiqot maqsadi mehnat sarfini kamaytirishga qaratilganda, faqat zaruriy izohlar majmuini o‘rganish maqsadga muvofiq. Demak, bu nuqtai nazardan, korpus razmetkasini chuqurlashtirish o‘zini oqlamaydi, razmetkani soddalashtirish yo‘lidan borish hamda qidiruv natijasini aniqlashtirishga e’tibor qaratish kerak

[1]Бабина О.И., Дюмин Н.Ю. Автоматизация лингвистической разметки корпуса текстов // http://helling100.narod.ru/pubs/AutomationBabinaDyumin.pdf

bo‘ladi.Minimallashtirish konsepsiyasiga asoslanadigan bo‘lsak, razmetka tizimiga faqat zaruriy axborotni kiritish maqsadga muvofiq. Boshqa tomondan, korpusda qo‘llanuvchi vositaning boshqa tadqiqiy korpusda qo‘llash mumkin bo‘lgan avtomatik razmetka metodologiyasi sifatida foydalanish samarali natija beradi. Shu konsepsiyadan kelib chiqqan holda, O.I.Babina, N.Yu.Dyuminlar lingvistik razmetka vositalarini tuzish prinsiplari sifatida quyidagilarni sanab o‘tishadi[1]:

1. Dasturiy ta’minot vositasi turli xil tizimlar o‘rtasidagi muvofiqlik muammosining oldini olishga yordam beradigan Unicode belgilar kodlash tizimini qo‘llab-quvvatlashi kerak, shu bilan kirill yoki lotin alifbosi bo‘lmagan alifbolar diakritikasi yoki alifbodan foydalanadigan tillarni tasvirlash imkoniyatini berishi kerak.

2. Matnlar to‘plami hamda tegishli lingvistik ma’lumot yagona ma’lumotlar bazasida saqlanishi kerak, bunda matn korpusi bilan ishlash uchun turli xil funksiyalarni amalga oshiradigan tizimning dasturiy komponentidan standartlashtirilgan kirish ta’minlanadi.

3. Korpusga ishlov berishga mo‘ljallangan ma’lum vositalar undan alohida bo‘lishi talab qilinadi. Shuning barobarida, tizimning umumiy universalligini ta’minlashishga erishish lozim. Ma’lumotning lingvistik bazadan boshqa tizimlarda ham takroriy qo‘llanishiga erishish mumkin.

4. Tizimning har bir komponenti alohida lingvistik vazifa bajaradi, tizimning modul tashkilotini ta’minlaydi.

5. Til materiali(matn)ning reprezentatsiyasi asosiy omil; barcha hosilaviy lingvistik ma’lumotlar (xususan, leksikon) matn korpusidagi pozitsiyalarga bog‘langan. Matn reprezentativligi prinsipi korpusdagi turli so‘zshakl, so‘z birikmasi leksik va grammatik kontekstiga erkin kirishga sharoit yaratadi.

6. Leksik birlikni matnga biriktirish esa omonim so‘zshakl va so‘z birikmalariga yoziladigan turli morfologik teglar majmuini ajratib olishga yo‘l ochadi.

[1]Ўша манба.

Matnning avtomatik razmetka majmui dasturiy vositalari quyidagilardan tashkil topadi:

1) korpusni boshqarish moduli (CorpusManager);
2) avtomatik morfologik razmetka moduli (AutoPOSTagger);
3) morfologik razmetkaning avtomatik korrektori (Corrector);
4) avtomatik sintaktik razmetka moduli (SynTagger).

Aytish joizki, ushbu tizim matnga har tomonlama ishlov berishga mo‘ljallangan. Shu bila birga, turli vazifalarni bajaruvchi vositalar alohida ishlab chiqiladi.

XANKOda an’anaviy sintaksis. Ma’lumki, an’anaviy yondoshuv asoslari XIX asrda rus tilshunoslari tadqiqotlarida mukammal ishlangan. Rus tili sintaksisining eng to‘liq tavsifini 1960 yillarning akademik grammatikasi deb hisoblash mumkin. Zamonaviy tasniflar hozirgi rus tili universitet darsliklarida biroz o‘zgarish, farq bilan aks ettirilgan[1]. Ushbu yondashuvning afzalliklari quyidagilar:

1) umumiylik va soddalik;
2) boshqa sintaktik yondashuvlar asosida (birinchi navbatda, struktur sintaksis), tadqiqot uchun materialni bilvosita izlash imkoniyatining mavjudligi;

An’anaviy sintaksisning kamchiliklari quyidagilarda ko‘rinadi:

1) sintaktik tuzilmalar tabiati to‘g‘risidagi zamonaviy g‘oyalar bilan nomuvofiqlik;
2) sintaktik birliklar tavsifi hamda sintaktik aloqani e’tiborsiz qoldirish;
3) tavsifdagi nomuvofiqlik, muqarrar qarama-qarshiliklar (predpozitsion guruhning yo‘qligi, turli bo‘lak ichidagi bo‘lakni aniq ajratib olmaslik);
4) avtomatik ishlov berishning murakkabligi.

Biroq yondashuvning ko‘rsatilgan afzallik/kamchiliklari ish natijasini oqlamaydi; aksincha, ular potensial foydalanuvchining taxminlarini yanada chigallashtiradi.

XANKO (Xelsinskiy annotirovannыy korpus qisqartmasi) yaratuvchilari oddiy foydalanuvchiga tushunarli bo‘lgan razmetka darajasini saqlab qolish uchun umumqabul qilingan nazariyalardan foydalanishni ma’qul ko‘rishgan. Quyida

[1] Валгина Н.С., Современный русский язык. Синтаксис. М.: Высшая школа, 2003.; Кустова Г.И., Мишина К.И., Федосеев В.А. Синтаксис современного русского языка. М., 2005.

XANKOni ishlab chiqilgan tamoyillarni keltirishni ma’qul ko‘rdik, zero, o‘zbek tili korpusi sintaktik teglar tizimini ishlab chiqishda hamda shunga o‘xshash holatlarda ushbu tamoyillarga tayanish mumkin.

1. Muayyan muammoni hal qilishda XANKO yaratuvchilari doim bu yoki boshqa sintaktik ma’lumot va uning axborot qayta ishlashni avtomatlashtirishda qanchalik muhimligiga e’tibor berishgan. Qo‘lda bajariladigan ishlarning kutilgan hajmi va natija qiymati ko‘pincha bir-biriga zid keladi: masalan, determinantni birlik sifatida ajratib ko‘rsatish qo‘lda qilinadigan ishlarning sezilarli darajada ko‘payishiga olib keladi (determinantni qidirishni avtomatlashtirish mumkin emas), ammo bu ishni ketma-ket bajarish qiyin bo‘ladi, chunki konsepsiyaning ko‘lami “determinant”ni turli matnda turlicha aniqlaydi.

2. Interfeysning qulayligini hisobga olish ham diqqat-e’tibordagi masala sanaladi. Sintaktik ma’lumotlar turli xil birlikka, jumladan, morfologik ma’lumotni o‘z ichiga olgan matn shakliga, ikki marta belgi qo‘yilgan hollarda kiritiladi.

Ushbu sintaktik tahlil tizimlari o‘zbek tili sintaktik tahlil tizimini yaratish uchun zaruriy tajriba maydoni bo‘lib xizmat qiladi. Yuqorida sanab o‘tilgan parser(sintaktik tahlil tizimlari)ni o‘rganar ekanmiz, sintaktik tahlil tizimi qanday tarkibiy qismlardan tashkil topishi, sintaktik tahlil teglarini ishlab chiqish uchun qanday lingvistik bilimlar kerak bo‘lishini kuzatdik. Xulosa sifatida aytish mumkinki, har bir tildagi sintaktik razmetka tizimini ishlab chiqish uchun o‘sha tilning sintaktik qurilishi modellashtirilishi talab etiladi. Modellashtirishdan keyingi bosqich sintaktik teglar tizimini tuzish, so‘nggi qadam esa matn til birliklariga sintaktik teglarni biriktirishdir.