



Aziz YULDASHEV,
ToshDO'TAU o'qituvchisi
E-Mail: yuldashevaziz@navoiy-uni.uz

ToshDO'TAU professori, f.f.d Sh.Hamroyeva taqriza asosida

LINGVISTIK VA EKSTRALINGVISTIK TEGTLASH XUSUSIDA

Аннотация

Ushbu maqolada lingvistik va ekstralingvistik teglash, uning til korpusi qidiruv tizimini ishlab chiqishdagi ahamiyati xususida fikr yuritilgan. Lingvistik teglash matn tahlili, modellarni o'qitish, lingvistik tadqiqotlarni amalga oshirishda muhim ahamiyat kasb etadi. Lingvistik teglash (izohlash) – bu matnga ma'lum bir lingvistik ma'lumot (izoh) qo'shish jarayonidir. Bu ma'lumot(izoh)lar so'z turkumi haqidagi izohdan tortib murakkab sintaktik qurilmalargacha bo'lishi mumkin. Korpus interfeysida lingvistik teglashni aks ettirish tadqiqotchining ish qulayligi va samaradorligini belgilovchi muhim jihatdir.

Kalir so'zlar: Til korpusi, lingvistik teglash, ekstralingvistik teglash, qidiruv tizimi, til korpusi qidiruv tizimi.

О ЛИНГВИСТИЧЕСКОЙ И ЭКСТРАЛИНГВИСТИЧЕСКОЙ МАРКИРОВКЕ

Аннотация

В данной статье обсуждается лингвистическое и экстралингвистическое тегирование, его значение при разработке поисковой системы языкового корпуса. Лингвистическое тегирование играет важную роль в анализе текста, обучении моделей и проведении лингвистических исследований. Лингвистическое тегирование (аннотирование) — это процесс добавления определенной лингвистической информации (аннотации) к тексту. Эта информация (аннотация) может варьироваться от аннотации о части речи до сложных синтаксических конструкций. Отображение лингвистического тегирования в интерфейсе корпуса является важным аспектом, определяющим удобство и эффективность работы исследователя.

Ключевые слова: Языковой корпус, лингвистическое тегирование, экстралингвистическое тегирование, поисковая система, поисковая система языкового корпуса.

ON LINGUISTIC AND EXTRALINGUISTIC LABELING

Annotation

This article discusses linguistic and extralinguistic tagging, and its significance in the development of a search engine for a language corpus. Linguistic tagging plays a crucial role in text analysis, model training, and conducting linguistic research. Linguistic tagging (annotation) is the process of adding specific linguistic information (annotations) to text. This information (annotation) can range from part-of-speech annotations to complex syntactic structures. The display of linguistic tagging within the corpus interface is a critical aspect that determines the convenience and efficiency of a researcher's work.

Key words: Language corpus, linguistic tagging, extralinguistic tagging, search engine, language corpus search engine.

Lingvistik teglash (izohlash) – bu matnga ma'lum bir lingvistik ma'lumot (izoh) qo'shish jarayonidir. Bu ma'lumot(izoh)lar so'z turkumi haqidagi izohdan tortib murakkab sintaktik qurilmalargacha bo'lishi mumkin. Lingvistik teglashning ahamiyati quyidagilarda namoyon bo'ladi:

Matn tahlili. Lingvistik teglash kompyuter/dastur(lar)ga matnning tuzilishi va ma'nosini tushunish imkonini beradi, bu esa tabiiy tillarni qayta ishlash (NLP) bo'yicha ko'plab vazifalarda ahamiyatli.

Modellarni o'qitish. Lingvistik teglangan ma'lumotlar NLPning turli xil vazifalarini bajara oladigan mashina o'rganish modellarni o'qitish uchun ishlatiladi. Masalan, mashina tarjimai, sentiment(kayfiyat) tahlili va nomlangan obyektlar(NER)ni aniqlash.

Lingvistik tadqiqotlar. Lingvistik teglash tilshunosga tilni o'rganish, qonuniyatlarni aniqlash va gipotezalarini tekshirishga yordam beradi[1].

Tabiiy tilni qayta ishlash sohasida lingvistik teglashning quyidagi turlari farqlanadi. Morfologik teglash so'z turkumlari, egalik, kelishik, shaxs/son, zamon, mayl, nisbat va boshqa morfologik belgilarni aniqlashni anglatadi. Sintaktik teglashda gapdagi so'zlar o'rtasidagi sintaktik munosabatlar – ega, kesim va to'ldiruvchi kabi bo'laklarni aniqlash tushuniladi. Semantik teglash so'z ma'nolari va ular orasidagi munosabatlarni (sinonim, antonim va giperonim) aniqlashdir. Diskurs teglash diskurs darajasidagi matn tuzilishini aniqlash jarayonidir. Masalan, mavzular, argumentlar va gaplar o'rtasidagi aloqalarni ajratish shunga misol bo'ladi. Anaforik teglashda matndagi koreferent aloqalar aniqlanadi. Masalan, olmoshlarning qaysi birlikka ishora etayotganligi belgilanadi. Prosodik teglashda birlik (so'z va gap

bo'lagi) urg'u va intonatsiyasi tavsiflanadi[2]. Ko'rinadiki, lingvistik teglash matnga uning morfologik belgisi, sintaktik xususiyatlarini, semantik o'ziga xosligi, qaysi diskursga oidligi, anaforik va hatto urg'u bilan bog'liq izoh qo'shish(yozish)dan iborat bo'ladi. Korpus matnlarida bu kabi ma'lumotlarning mavjud bo'lishi korpus qidiruv tizimini ishlab chiqish uchun juda muhim ahamiyat kasb etadi.

V.Zaxarovning fikriga ko'ra, bularning barchasi quyidagi tamoyillarga muvofiq amalga oshiriladi: teglash sxemasining tavsifi (asoslanishi); lingvistik tushunchalarning umumqabul qilingan tizimi; foydalanuvchi uchun tanish bo'lgan tahlil sxemasi; parametrlarni kiritishning asosiligi; nazariy jihatdan betaraf (an'anaviy) teglash sxemasi; xalqaro standartlarga rioya qilish[3].

Morfologik teglash ko'pincha "part-of-speech tagging" (POS-tagging) deb ataladi, so'z turkumini aniqlash degan ma'noni anglatadi. Aslida, morfologik teglar faqat so'z turkumini ko'rsatib qolmay, balki unga xos bo'lgan grammatik kategoriyalarni ham o'z ichiga oladi. Bu teglash usuli eng muhim qadamdir: birinchidan, katta matnlar bazasining ko'p qismi aynan morfologik teglangan; ikkinchidan, morfologik tahlil keyingi sintaktik va semantik tahlillar uchun asos bo'lib xizmat qiladi; uchinchidan, kompyuter morfologiyasi sohasidagi yutuqlar katta hajmdagi matnlarni yuqori aniqlik bilan avtomatik ravishda teglash imkonini beradi.

Teglash ma'lumotlari muayyan tartibda taqdim etiladi va lemma, so'z turkumi va grammatik kategoriyalarni o'z ichiga oladi. 1980 yilda Braun korpusining lingvistik teglangan versiyasi yaratildi, unda so'z shakllari lemmatizatsiya qilindi, ularning sintaktik vazifalari belgilandi.

Sintaktik teglash morfologik tahlil ma'lumotlari asosida bajariladigan parsing natijasidir. Teglashning bu turi leksik birliklar o'rtasidagi sintaktik aloqalar va turli sintaktik konstruksiyalarni (masalan, ergash gap, fe'lli birikma va boshqalar) tasvirlaydi[3]. Morfologik teglashdan farqli o'laroq, sintaktik qurilma va sintaktik munosabatlarni taqdim etish usullari bir xil emas. Sintaktik qoidalarning quyidagi xilma-xilligi kuzatiladi: so'zlar orasidagi sintaktik munosabat turlari (bog'lanish grammatikasi); birikma a'zolari; qurilish sxemalari grammatikasi; gap bo'laklari haqidagi an'anaviy sintaktik ta'limotlar; funksional grammatika.

Ko'plab tillarni lingvistik teglash uchun maxsus vositalar mavjud bo'lib, ularning ayrim xususiyatlari universallik kasb etadi. Lingvistik teglash uchun ko'plab vositalar mavjud: avtomatlashtirilgan va qo'lda amalga oshiriladigan bunday vositalar sifarsiga Stanford CoreNLP[4], spaCy[5] kabi vositalarni kiritish mumkin.

Demak, lingvistik teglash – bu tabiiy tillarni qayta ishlash va lingvistik tadqiqotlar uchun muhim vositadir.

Korpus interfeysida lingvistik teglashni aks ettirish tadqiqotchining ish qulayligi va samaradorligini belgilovchi muhim jihatdir. Bu ma'lumotni vizuallashtirishning bir necha usullari mavjud, ularning har biri o'z afzallik va kamchiliklariga ega. Asosiy aks ettirishning quyidagi usullari farqlanadi:

Qatlamli aks ettirish. Bu usul foydalanuvchiga teglashning turli qatlamlari (masalan, morfologiya, sintaksis, semantika) o'rtasida o'tish imkonini beradi. Bu ma'lumotning aniq ajratilishini ta'minlaydi va teglashning aniq turiga e'tibor qaratishga imkon beradi. Masalan, Rus tili milliy korpusi (NKRYA) lingvistik annotatsiyaning turli darajalarini ko'rish imkonini beradi.

Parallel aks ettirish. Teglashning bu turi matn yonida, ko'pincha ustun yoki annotatsiya shaklida aks ettiriladi. Bu bir vaqtning o'zida bir necha tahlil darajalarini ko'rishga imkon beradi, bu ular o'rtasidagi munosabatlarni o'rganish uchun qulay. Biroq katta hajmdagi teglashda bu interfeysni tig'iz qilib qo'yishi mumkin.

O'rnatilgan teglash (annotatsiya). Teglash maxsus belgilar yoki teglardan foydalangan holda to'g'ridan-to'g'ri matnga o'rnatiladi. Bu teglashni kontekstda ko'rishga imkon beradi, ammo murakkab (mukammal) teglashda o'qishni qiyinlashtirishi mumkin.

Sintaktik qurilmalarni vizuallashtirish. Shajara daraxti yoki graflar so'zlar o'rtasidagi sintaktik munosabatlarni aks

ettiradi. Bu gap tuzilishini vizual ravishda tasavvur qilishga yordam beradi va sintaksis tahlilini osonlashtiradi.

Kontekstli ko'rinish. Kursorni so'z yoki ibora ustiga olib borganda, batafsil lingvistik ma'lumotga ega qalqib chiquvchi oyna paydo bo'ladi. Bunday yondashuv foydalanuvchiga interfeysni tig'iz qilmasdan, talab bo'yicha ma'lumot olish imkonini beradi.

Lingvistik teglash natijasining interfeysda aks ettirishning asosiy jihatlari quyidagilar: Sozlanuvchanlik: foydalanuvchi teglashni aks ettirishni sozlash, uning qiziqishiga mos qatlamlar va mukammallik darajasini tanlash imkoniga ega bo'lishi kerak. Qidiruv va filtrlash: qidiruv va filtrlash funksiyalari muayyan teglarga ega bo'lgan matnning kerakli qismlarini tez topish imkonini beradi. Ma'lumotlarni eksport qilish: kyevingi tahlil uchun belgilangan ma'lumotlarni turli formatlarda (masalan, XML, CSV) eksport qilish imkoniyatini beradi. Interaktivlik. Interfeys interaktiv bo'lishi maqsadga muvofiq, bu foydalanuvchiga korpus bo'ylab oson harakatlanish va matnga qo'yilganlar bilan bog'lanish imkonini beradi. Tushunarliklik: teglashda ishlatiladigan teglar tizimi aniq va tushunarli bo'lishi kerak.

Yuqorida keltirilganlar asosida teglangan korpuslar sifatida quyidagilarni misol keltirish mumkin:

Rus tili milliy korpusi (NKRYA): lingvistik belgilangan matnlar bilan ishlash uchun kuchli va moslashuvchan interfeysni ta'minlaydi.

Stanford CoreNLP: sintaktik daraxtlarni vizuallashtirishni o'z ichiga olgan tabiiy tillarni qayta ishlash uchun vositalar to'plamini taqdim etadi.

SpaCy: matndagi bog'liqliklarni vizuallashtirish imkoniyatini beradi.

Til korpusida teglash natijasini aks ettirish usulini tanlash tadqiqot maqsadlari, teglash turi va foydalanuvchining qiziqishiga bog'liq. Korpus bilan ishlashni osonlashtiradigan intuitiv va qulay interfeys yaratish muhim.

Ekstralingvistik teglash va uni interfeys qidiruv tizimida aks ettirish xususida

Ekstralingvistik teglash (yoki myetarazmetka) matnning lingvistik tuzilishidan tashqariga chiqib, u bilan bog'liq (nolisoniy) tashqi omillar haqidagi ma'lumotlarni o'z ichiga oladi. U matnni yaxshiroq tushunish va talqin qilishga yordam beradigan kontekstni taqdim etadi. Ekstralingvistik teglar quyidagi asosiy jihatlarni qamrab oladi (Qarang: 1-jadval).

Ekstralingvistik teg	Mazmuni
Muallif haqida ma'lumot	Muallifning ismi Jinsi Yoshi Ijtimoiy maqomi Ta'limi
Matn haqida ma'lumot	Nomi Janri (masalan, maqola, roman, yangilik) Nashr etilgan sanasi Nashr etilgan joyi Mavzusi Maqsadli auditoriyasi
Ijtimoiy-madaniy kontekst	Tarixiy davr Madaniy me'yorlar Ijtimoiy voqealar
Texnik ma'lumot	Fayl formati Kodlash Qayta ishlangan sanasi Teglash ijrochilari
Paralingvistik omillar	Mimika, imo-ishoralar, pauzalar, ohang, kulish, yig'lash

1-jadval. Ekstralingvistik teg turlari va mazmuni

Ekstralingvistik tegning ahamiyati quyidagilarda ko'rinadi: Matnni to'liq tahlil qilish. Ekstralingvistik teglash tadqiqotchi va tahlilchilarga matn yaratilgan kontekstni hisobga olish imkonini beradi, bu uning ma'nosini chuqurroq tushunishga olib keladi. Kontekstning tilga ta'sirini o'rganish. Ekstralingvistik (nolisoniy) teglar ijtimoiy, madaniy va tarixiy omillarning tildan foydalanishga ta'sirini aniqlashga yordam beradi. NLP tizimlarini takomillashtirish. Ma'lumot(lar)ga ekstralingvistik ma'lumot qo'shish tabiiy tilni qayta ishlash modellarining ishlashini yaxshilashi, ularni yanada aniq va kontekstual xabardor qilishi mumkin. Korpus lingvistikasi. Ekstralingvistik (nolisoniy) teglash korpus lingvistikasida muhim rol o'ynaydi, bu tadqiqotchiga til

o'zgarish va variatsiyasini turli ijtimoiy va madaniy omillarga qarab o'rganish imkonini beradi.

Ekstralingvistik teglashdan quyidagi o'rinlarda foydalanish mumkin: Ijtimoiy tarmoqlarni tahlil qilish. Ijtimoiy tarmoqlar ma'lumotlarini teglash foydalanuvchi profili, ularning joylashgan joyi va nashr etish vaqti haqidagi ma'lumotlarni o'z ichiga olishi mumkin, bu tendensiyalar va jamoatchilik fikrini tahlil qilish imkonini beradi. Tarixiy tadqiqotlar. Tarixiy matnlarni teglash muallif, tarixiy voqealar va ijtimoiy sharoitlar haqidagi ma'lumotlarni o'z ichiga olishi mumkin, bu tadqiqotchilarga tilning tarixda qanday ishlatilganini tushunishga yordam beradi. Videoreklamalarni tahlil qilish. Ekstralingvistik

tyeglash videodagi qahramonlarning mimikasi, imo-ishoralari haqidagi ma'lumotlarni o'z ichiga olishi mumkin.

Ekstralingvistik teglash matnini to'liq va chuqurroq tushunishni ta'minlab, lingvistik tahlil imkoniyatlarini kengaytiradi. Korpus interfeysida ekstralingvistik teglashni aks ettirish kontekst va ma'lumotlarni tashkil etishga alohida e'tibor talab qiladi. Quйда bu masalada bir nechta yondashuv va tavsiyalar keltiriladi:

1. Hujjat darajasidagi metama'lumotlar: Alohida panel. Matn haqidagi barcha ekstralingvistik ma'lumotlar taqdim etiladigan alohida panel yoki yorliq yaratiladi. Bu o'z ichiga quyidagilarni olishi mumkin: Muallif haqida ma'lumot (tarjimai hol, demografik ma'lumotlar). Nashr haqida ma'lumot (sana, manba, janr). Ijtimoiy-madaniy kontekst (tarixiy voqealar, madaniy me'yorlar). Qalqib chiquvchi oyna. Kursorni hujjat nomi yoki muallif ismi ustiga olib borganda, qisqa ma'lumot bilan qalqib chiquvchi teglar aks ettiriladi. Filtrlar va qidiruv. Foydalanuvchilarga hujjatlarni turli ekstralingvistik parametrlar (masalan, muallif, nashr sanasi, janr bo'yicha) bo'yicha filtrlash va qidirish imkoniyatini beruvchi filtr ishlab chiqiladi[6].

2. Fragment darajasidagi kontekst ma'lumotlari: Annotatsiyalar. Matnning aniq fragmentlariga ekstralingvistik ma'lumot qo'shish uchun annotatsiyalardan foydalaniladi. Masalan, tarixiy voqea haqida eslatma bo'lgan fragmentni tegishli maqola yoki hujjatga havola bilan annotatsiyalash mumkin. Vizualizatsiya. Matn va ekstralingvistik ma'lumotlar o'rtasidagi aloqalarni aks ettirish uchun vizualizatsiyadan foydalanish samarali natija beradi. Masalan, matn(lar) nashr etilgan sana(lar) va asosiy tarixiy voqealar aks ettiriladigan vaqt shkalasini yaratish mumkin.

3. Interaktivlik va sozlanuvchanlik: Interaktiv unsur(element)lar. Interfeysni interaktiv qilish ahamiyatga ega,

shunda foydalanuvchi qo'shimcha ma'lumotni olishi osonlashadi. Masalan, tegishli hujjat yoki veb-sahifalarga o'tish uchun giperhavolalardan foydalanish mumkin. Sozlanuvchan aks ettirish. Foydalanuvchiga o'zi qiziqqan parametrlarni tanlab, ekstralingvistik ma'lumotlarni aks ettirishni sozlash imkoniyatini berish lozim (Bizning ishimizning amaliy natijasi aynan shunga qaratiladi).

4. Texnik jihatlar: Standartlashtirish. Ekstralingvistik ma'lumotlarni saqlash va almashish uchun standartlashtirilgan formatlardan (masalan, XML, JSON) foydalanish maqsadga muvofiq. Ma'lumotlar bazalari. Katta hajmdagi ekstralingvistik ma'lumotlarni saqlash va boshqarish uchun ma'lumotlar bazalaridan foydalaniladi[7].

Ekstralingvistik teglash natijasida korpus qidiruvining ishlab chiqilishi natijasida podkorpus (ichki korpus) hosil qilish mumkin. Bu amaliyot ayrim korpuslarda yo'lga qo'yilgan (Podkorpus masalasi ishning alohida bo'limida tahlil qilinadi). Masalan, Rus tili milliy korpusi (NKRYA)[8] har bir matn haqida batafsil, jumladan, muallif, nashr sanasi va janri ma'lumot beradi. Foydalanuvchi matnlarni turli parametrlar bo'yicha filtrlashi va qidirishi mumkin. Yoki milliy korpuslar[9] tarkibidagi tarixiy korpuslar[10] ni hosil qilishda ham davrlashtirish muammosi ekstralingvistik teglar yordamida amalga oshiriladi. Tarixiy korpuslar ko'pincha hujjat(asar) yaratilgan sana, muallif haqida ma'lumot va tarixiy kontekst kabi batafsil ekstralingvistik ma'lumotlarni o'z ichiga oladi. Bunda muallif, hujjatlar va tarixiy voqealar o'rtasidagi aloqalarni aks ettiruvchi graflarni vizualizatsiya qilish ham muhim ahamiyat kasb etadi[11]. Masalan, Rus tili milliy korpusida metama'lumotlar asosida til birliklari davriy xususiyatlarini kuzatishda vizualizatsiyadan unumli foydalanilgan (Qarang: 2-rasm)[12].



2-rasm. Rus tili milliy korpusida ekstralingvistik teglash natijasi qidiruv vizualizatsiyasi

Xulosa sifatida aytish mumkinki, lingvistik va ekstralingvistik ma'lumotlar va teglar korpus qidiruv tizimini ishlab chiqishda muhim ahamiyat kasb etadi.

ADABIYOTLAR

1. <http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127221;>
2. https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D0%BF%D1%83%D1%81_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2
3. <https://studfile.net/preview/1848098/>
4. <https://stanfordnlp.github.io/CoreNLP/>
5. <https://spacy.io/>
6. Frensis U.N. Проблемы формироваия и машинного представлениа болшого корпуса текстов / U.N. Frensis // Novoye v zarubejnoj lingvistike. Выпуск Проблемы и методы лексикогрфии. – М.: Progress. – 1983. – S. 334 – 335.
7. Vensov A.V., Grudeva Ye.V., Kasevich V.B., Yagunova Ye.V. Natsionalnyy korpus russkogo literaturnogo yazyka: nekotorye rezultaty, prilozheniya i zadachi / Vensov A.V., Grudeva Ye.V., Kasevich V.B., Yagunova Ye.V. // Nauchno-texnicheskaya informatsiya. – Ser. 2. – № 6. – S. 35–36.
8. <https://ruscorpora.ru/>; The Uppsala Russian Corpus: [elektronnyy resurs]: <http://www.slaviska.uu.se/korpus.htm> (data obraeniya: 05.12.2024); The Lancaster-Oslo/Bergen Corpus of British English, LOB: [Electronic resource]: <http://clu.uni.no/icame/manuals/LOB/INDEX.HTM> (accessed: 04.12.2024).
9. <https://www.english-corpora.org/bnc/>; Brown Corpus: [elektronnyy resurs]: <http://clu.uni.no/icame/brown/bcm.html#bc3> (data obraeniya: 03.03.2025)
10. <https://ruscorpora.ru/search?search=CgQyAggP>
11. Sharov S.A. Predstavitelnyy korpus russkogo yazyka v kontekste mirovogo opyta / S.A. Sharov // Nauchno-texnicheskaya informatsiya. – Ser. 2. – № 6. – S. 12–16.
12. <https://ruscorpora.ru/results?search=Cs8BEpwBCpkBChMKCWRpc2FtYm1vZBIGCgRtYWluChcKB2Rpc3Rtb2QSDAoKd2l0aF96ZXJvcXpChEKA2xleBIKCGjQtNC%2B0LvRjwoKCGRmb3JtEgIKAAoLCgVncmFtbRICCGAKCQoDc2VtEgIKAAoVCgdZw0tbW9kEgoKCHNlbXxzZW14CgwKBnN5bnRheBICCGAKCwoFZmxhZ3MSAgAKicKCAgAEaAoYMiAKEAUgAEAFSGdjcmVhdGVkagQwLjk1eACyAQGgAQEYAggBOgEF>