



O‘ZBEK TILI KORPUSIDA REGULAR IFODA YORDAMIDA LINGVISTIK QIDIRUVNI AMALGA OSHIRISHNING AFZALLIKLARI

Yuldashev Aziz Uyg‘un o‘g‘li
ToshDO‘TAU tadqiqotchisi
E-mail: yuldashevaziz@navoiy-uni.uz

Annotatsiya. Mazkur maqolada til korpusida lingvistik qidiruvni amalga oshirish uchun regular ifodadan foydalanish va uning natijalari bo‘yicha tahlil olib borildi. Korpus matnlarida qidiruvni amalga oshirishda regular ifodalardan (regex) foydalanish qidiruv imkoniyatlarini sezilarli darajada kengaytirishga, uni yanada moslashuvchan va aniq qilishga imkon beradi. Regular ifodalar NLPda juda muhim rol o‘ynaydi va matnni qayta ishlashning ko‘plab vazifalarini bajarishga yordam beradi.

Kalit so‘zlar: regular ifoda, lingvistik qidiruv, agglyutinatativ til, konkordans.

Abstract. This article analyzes the use of regular expressions for linguistic search in the language corpus and its results. The use of regular expressions (regex) in searching corpus texts allows for a significant expansion of search capabilities, making it more flexible and precise. Regular expressions play a very important role in NLP and help to perform many tasks of text processing.

Keywords: regular expression, linguistic search, agglutinative language, concordance.

Аннотация. В данной статье был проведен анализ использования регулярного выражения для осуществления лингвистического поиска в языковом корпусе и его результатов. Использование регулярных выражений (regex) при выполнении поиска в корпусных текстах позволяет значительно расширить возможности поиска, сделать его более гибким и точным. Регулярные выражения играют очень важную роль в NLP и помогают выполнять многие задачи обработки текста.

Ключевые слова: регулярное выражение, лингвистический поиск, агглютинативный язык, конкорданс.

O‘zbek tili korpusida qidiruv uchun regular ifodalardan foydalanish lingvistik tadqiqotlar uchun keng imkoniyatlar ochadi. O‘zbek tilining xususiyatlaridan kelib chiqib regular ifodalarni qo‘llash o‘ziga xos xususiyatlarga ega bo‘ladi. **Regular ifoda (regex)** – bu matndagi belgilar ketma-ketligi uchun moslashuvchan qidiruv qolipi bo‘lib, u aynan ko‘rsatilgan naqshga mos keluvchi istalgan matn qismini topishga xizmat qiladi. Oddiy qilib aytganda, regular ifoda yordamida *“shu va shu shartlarga mos keladigan so‘z yoki gaplarni topib bering”* deb so‘rov yuborish mumkin. Masalan, $^{\wedge}[A-Za-z]^{\$}$ degan regex barcha faqat



lotin harflaridan tashkil topgan soʻzlarni topadi, yoki **lar\$** degan oddiy ifoda soʻz oxirida “*lar*” ketma-ketligi bilan tugaydigan elementlarni aniqlaydi. Korpusda regular ifodalar orqali qidiruv bu oddiy matn boʻyicha qidiruvdan kengroq imkoniyatlar beruvchi qudratli vosita, chunki regex yordamida bitta ifoda bilan cheksiz koʻp shakldagi soʻzlarni qamrab olish mumkin¹: foydalanuvchi oldindan roʻyxatini tuzib boʻlmaydigan barcha potensial moslamalarni bitta umumiy qolip orqali izlashi mumkin.

Agglyutinatív tillarda, jumladan, oʻzbek tilida soʻz tarkibiga turli affikslar qoʻshilishi orqali koʻp maʼno va grammatik shakl hosil qilinadi. Regular ifodalar bu affikslarni qidirishda juda qoʻl keladi. Misol uchun, oʻzbek tilida koʻplik qoʻshimchasi “*-lar*” boʻlib, **\w+lar\b** kabi regex yozuvi matnda barcha “*-lar*” qoʻshimchasi bilan tugagan soʻzlarni topadi. Bu yerda **\w+** – istalgan uzunlikdagi soʻz tanasi, **lar** – shu harflar ketma-ketligi, **\b** – soʻz chegarasi. Shu tariqa, regex yordamida korpusdan koʻplik shakllarning barchasini avtomatik terib chiqish, ularning statistik tarqalishini oʻrganish mumkin. Shuningdek, feʼlning maʼlum zamon yoki nisbat affikslarini qidirish orqali muayyan grammatik hodisaning uchrash tezligini tekshirish mumkin. Masalan, oʻzbek tilida hozirgi davomli zamon feʼli “*-yapti*” qoʻshimchasini oladi. Regex yordamida **\w+yapti\b** soʻrovi korpusda hozirgi davomli zamonda qoʻllangan barcha feʼllarni ajratib beradi.

Korpusda soʻzlar soʻz turkumlari boʻyicha teglangan boʻlsa, regular ifodalar yordamida maʼlum turkumdagi soʻzlarni yoki turkumlar ketma-ketligini qidirish mumkin. Masalan, agar korpusda har bir tokenning yonida uning turkumi belgilangan boʻlsa (kitob/N yoki yaxshi/Adj tarzida), unda **\b\w+/N\b** ifodasi barcha otlarni topib beradi, yoki **Adj\s+\w+/N** ifodasi sifat + ot tartibida kelgan birikmalarni topishi mumkin (“*qizil gul*”, “*yangi loyiha*” kabi). Bu usul yordamida matndan sintagmatik tuzilmalarni (sifat + ot birikmalari, ravish + feʼl birikmalari) avtomatik ajratish va ularni tahlil qilish imkoni tugʻiladi. Agar korpusda teglar boʻlmasa ham, regex yordamida maʼlum morfologik qolip asosida soʻz turkumini taxminan aniqlash mumkin. Masalan, oʻzbek tilida otlarning aksariyati unli bilan tugasa (*-a*, *-o* kabi) va sifatlar *-li*, *-siz* kabi qoʻshimchalar bilan tugashi mumkin. Shunday qoida asosida **\w+li\b** ifodasi sifatlarni qidirishda yordam berishi mumkin, garchi bunday qidiruv aniqligi annotatsiyalangan korpusdagidek yuqori boʻlmasa-da, umumiy tendensiyalarni chiqarib berishi mumkin.

Oʻzbek tilida soʻz yasovchi va shakl yasovchi qoʻshimchalar ketma-ketligi maʼlum morfologik qoliplarni hosil qiladi. Regular ifodalar shu qoliplarni ham aniqlash uchun ishlatiladi. Masalan, otlarda egalik qoʻshimchasi va kelishik qoʻshimchasi birikib kelsa, soʻz oxirida “*-ning*” (birlik egalik + kelishik) yoki “*-lari*” (koʻplik egalik) koʻrinishlari paydo boʻladi. **\w+ning\b** ifodasi barcha egalik kelishigidagi soʻzlarni topadi, **\w+lari\b** esa koʻplikdagi egalik formalarini ajratadi. Shu tariqa, korpusdan aniq bir morfologik koʻrinishni qidirib, u qanchalik tez-tez ishlatilishini bilib olish, boshqa shakllar bilan qiyoslash

¹ Smith, G. (2003). Searching for morphological structure with regular expressions. Tiger Projektbericht, Univ. Potsdam.
www.innovativepublication.uz



imkoniyati mavjud. Fe'llarning tuslanishidagi qoliplar: masalan, buyruq maylidagi fe'llar ko'pincha **-(V)ng** qo'shimchasi bilan tugaydi (*“boring”, “oling”*). Regex yordamida **\w+ng\b** degan ifoda barcha ana shunday buyruq shaklidagi fe'llarni topib, buyruq ohangining korpusdagi ulushini aniqlashga xizmat qiladi.

Regular ifoda bo'yicha qidiruv natijalari odatda ro'yxat shaklida chiqadi. Bu ro'yxatda qaysi so'zlar yoki iboralar ifodaga mos kelgani va ular matnning qaysi qismida uchragani ko'rsatiladi. Korpus lingvistikasi amaliyotida bunday natijalarni konkordansiya (concordance) ko'rinishida taqdim etish odatiy holdir. Konkordansiyada har bir topilgan misol alohida qatorda, uning chap va o'ng tomonidagi matn konteksti bilan birga ko'rsatiladi. Bu usul topilgan til hodisasini kontekstda tahlil qilish imkonini beradi. Misol uchun, regex yordamida topilgan *“gan edi”* shaklidagi o'tgan zamon fe'llari konkordansda chiqarilsa, tadqiqotchi ularning atrofidagi so'zlar bilan birgalikda qanday ma'no ifodalashini (masalan, hikoya ohangini yoki kinoyani) ko'ra oladi. Natijalarning bunday tavsifi lingvistik xulosalar chiqarish uchun muhim: faqat qatorlab sanab chiqishdan ko'ra, konkordansdagi kontekst ular qanday ishlatilishini tushunishga yordam beradi.

Xulosa qilib aytganda, regular ifodalar yordamida til korpusidan lingvistik hodisalarni tezkor va aniq aniqlash imkoniyatlari mavjud. Ayniqsa, agglutinativ til bo'lgan o'zbek tilida affiks, qo'shimcha, morfologik shakllarni avtomatik tarzda aniqlash juda qulay. Konkordans shaklida natijalarni chiqarish va tahlil qilish orqali muayyan lingvistik holatlarning qo'llanilish kontekstlarini o'rganish imkoniyati mavjud. Bu usul yordamida korpusda murakkab grammatik va morfologik tuzilmalarning statistik tahlilini ham o'tkazish mumkin.

Foydalanilgan adabiyotlar ro'yxati

1. http://www.machinelearning.ru/wiki/images/5/50/Korolev_RegExp.pdf
2. <https://learn.javascript.ru/regular-expressions>
3. https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D1%8B%D0%B5_%D0%B2%D1%8B%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D1%8F
4. Smith, G. (2003). Searching for morphological structure with regular expressions. Tiger Projektbericht, Univ. Potsdam.
5. Zakharov, K. Регулярные выражения и их применение в работе с открытыми данными.
6. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272 с.
7. Введение в регулярные выражения // <https://www.williamspublishing.com/PDF/5-8459-0937-6/part.pdf>
8. Фридл Дж. Регулярные выражения, 3-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2008. – 608 с. (Mastering Regular Expressions Jeffrey E.F. Friedl).