SODDA GAPLARDAGI MAZMUNIY MOSLIK, O‘XSHASHLIK VA SHAKLIY FARQLILIKNING LEKSIK VOSITALARI HAMDA ULARNING LINGVISTIK MODELLARI

Muratbekova Shoira Dilshodbek qizi
Alisher Navoiy nomidagi Toshkent davlat
o‘zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi mutaxassisligi
2-kurs magistranti

Annotatsiya. O‘zbekistonda ilmiy-tadqiqotning rivojlanishi, original ilmiy matnlarning paydo bo‘lishi, muayyan lingvistik nazorat (ko‘chirmakashliklarning oldini olish uchun) tartibini ishlab chiqishni taqozo qiladi. Buning uchun elektron shaklda mavjud ilmiy matnlarning mazmuniy o‘xshashlik darajasini aniqlaydigan dasturiy ta’minot yaratilishi zarur.

Kalit so‘zlar: *model, lingvistik model, Levenshteyn masofasi, Shingle algoritmi, Jakkard algoritmi.*

Abstract. The development of research in Uzbekistan, the emergence of original scientific texts requires the development of a certain linguistic control (to prevent duplication). For this, it is necessary to create software that determines the level of similarity of the content of scientific texts available in electronic form.

Key words: *model, linguistic model, Levenstein distance, Shingle's algorithm, Jaccard's algorithm.*

Аннотация. Развитие научных исследований в Узбекистане, появление оригинальных научных текстов требует разработки специальной процедуры лингвистического контроля (для предотвращения плагиата). Для этого необходимо создать программное обеспечение, определяющее степень сходства содержания научных текстов, имеющихся в электронном виде.

Ключевые слова: *модель, лингвистическая модель, расстояние Левенштейна, алгоритм Шингла, алгоритм Жаккара.*

Tilshunoslikda ham model tushuncha sifatida lingvist tomonidan tilning sun’iy ko‘rishdagi yangidan tartiblashtirilishi, ya’ni odatdagi ko‘rnishidan ko‘ra anchagina soddalashtirilgan shaklga keltirilishi. Bu o‘rinda lingvistik maqsadlarda tilning asl nusxasiga tegishli jihatlarni, so‘zlar, gap qoliplari (xatti-harakatlarini takrorlaydi)ga taqlid qiladi.

Tilshunoslikda modellarning ko‘plab ta’riflari mavjud:

- model – har qanday matn birliklarining (so‘z, gaplar) turi, qolipi (til qolipi);
- model – til obyektlarini tavsiflash uchun belgilar, sxemalardir;

- model– qat’iy metatilga ega bo‘lgan tuzilishning rasmiylashtirilgan nazariyasi.

Tilshunoslikda modellashtirishning asosiy maqsadi shaxsning integral lingvistik qobiliyatini modellashtirish va bu orqali uning ishlarini tez, oson va qulay amalga oshirishga sharoit yaratish. Lingvistik model tushunchasi strukturaviy tilshunoslikda vujudga kelgan, ammo 60–70-yillarda ilmiy foydalanishga kirgan. Zamonaviy tilshunoslikda “model” atamasining mazmuni asosan “nazariya” atamasi nomi ostida Yelmslev[1] tomonidan ilgari surilgan. Model faqat yetarlicha aniq ifodalangan va yetarlicha rasmiylashtirilgan nazariya nomiga loyiq deb hisoblanadi[1]. Har bir model ideal holda, kompyuterda amalga oshirilishi kerak.

Tilshunoslikda “lingvistik model” tushunchasi struktur lingvistikaning ta’siri natijasida paydo bo‘lgan va keyinchalik kompyuter lingvistikasini paydo bo‘lishiga olib kelgan[2]. Bundan ma’lum bo‘ladiki, lingvistik model hodisa sifatida struktur tilshunoslik va kompyuter lingvistikalarini bog‘lab turadi. Modelni qurish nafaqat lingvistik hodisalarni aks ettirish vositalaridan biri, balki til haqidagi bilimlarning haqiqatini tekshirishning obyektiv amaliy mezoni sifatida ham alohida ahamiyatli ishlardandir. Til o‘rganishning boshqa usullari bilan birlikda modellashtirish nutq faoliyatining yashirin mexanizmlari, uning nisbatan ibtidoiy modellardan tilning mohiyatini to‘liqroq ochib beradigan yanada mazmunli modellarga o‘tishi haqidagi bilimlarni chuqurlashtirish vositasi sifatida ishlaydi, desak mubolag‘a bo‘lmasa kerak.

Tilshunoslikda modellashtirish tizim sifatida o‘z prinsipiga ega va uning ba’zi quyi tizimlari boshqalarni modellashtiradi, masalan, yozma nutq tizimi og‘zaki nutq modelidir; yozma tilda biz bir nechta modellar (bosma, qo‘lda yozilgan) bilan ishlaymiz; ifoda rejasi kontent rejasi modeli.

Modellashtirish usuli odatda imo-ishora tizimlariga asoslanadi, lekin tilning o‘zi ham ishoralar tizimi, ya’ni so‘zlarni so‘zlar bilan modellashtirishdir. Har qanday model, shu jumladan lingvistik model ham rasmiy bo‘lishi kerak. Model, agar boshlang‘ich ma’lumot hamda matnlarni va ular bilan ishlash qoidalarini (yangi obyektlar va fikrlarni shakllantirish yoki joylashtirish qoidalari) aniq va ixcham ko‘rsata olsa, albatta ko‘zlangan maqsad hosil bo‘ladi hamda bu ko‘rinishida rasmiylik yuzaga keladi. Ideal holda, har qanday rasmiy model matematik tizim hisoblanadi. Shuning uchun, ma’lum ma’noda, rasmiyatchilik tushunchasi matematika, aniqlik yoki noaniqlik tushunchasiga teng bo‘lishi talab qilinadi. Rasmiylik, aniqlik, noaniqlik – bular nazariya taqdim etilgan tilning xususiyatlaridir. O‘z-o‘zidan bu xususiyat rasmiy nazariya bashoratlarining obyektiv eksperimental ma’lumotlar bilan mos kelishini ta’minlamaydi. Nazariyaning to‘g‘riligi uni tasdiqlash yoki rad etishga qodir bo‘lgan aniq tajribalarni o‘rnatishga imkon beradi, ammo nazariyaning aniqligi va haqiqati o‘rtasida zarur mantiqiy bog‘liqlik bor yoki yo‘q bo‘lishi amaliy tajribalarsiz xulosa berish mantiqan xato bo‘lgan hodisa. Rasmiy model eksperimental ma’lumotlar bilan u yoki bu talqin orqali bog‘lanadi. Modelni talqin qilish – bu model obyektlari (ramzlari) o‘rniga ma’lum bir mavzu sohasidagi

obyektlarni, masalan, tilni almashtirishning ehtimollik yoki qat’iy qoidalarini ko‘rsatishni anglatishi bilan xarakterlanadi.

Ishimizning dastlabki qismlarida ko‘chirmakashlik va uning oldini olish bilan bog‘liq mavjud ahvolni, ya’ni mualliflik huquqining himoya qilinishining huquqiy asoslari yetarli bo‘lsa-da, bu borada ilmiy yondashuvlarning ancha ortda qolayotganligini ta’kidlagan edik. Falsafada miqdor o‘zgarishi sifat o‘zgarishiga ta’sir qilishi qonuniyat sifatida qaralgani holda, O‘zbekiston Respublikasi Mustaqilligiga o‘ttiz yildan oshganiga qaramay, tadqiqotlarni amalga oshirish masalasida, ayniqsa tilshunoslikda jahon tilshunosligi bilan bellasha oladigan darajada dolzarb muammolar yechimi ko‘zga tashlanmayapti. Albatta, qonun kuchli, biroq inson ehtiyojlari undanda kuchli, degan qarashga ko‘ra, o‘zganing ishini noqonuniy o‘zlashtirish, ilmiy odobni chetlab o‘tish jiddiy muammo bo‘lib qolmoqda. Yuqorida ta’kidlaganimizdek qonunchilik nimani qilish mumkin, nimani qilish mumkin emas, taqiqlangan ishni qilib qo‘lga tushganda jazolanishini amalga oshira oladi. Biroq mualliflik huquqini buzilishi ilmiy asoslab berilmasa, chuqur tahlil metod va vositalariga ega bo‘lmasa, samarali natija kutish o‘rinsiz va imkonsiz bo‘lib qolaveradi. Bugungi kunda rivojlangan mamlakatlarda ham amalga oshirilayotgan ilmiy-tadqiqot ishlarining original matnligini aniqlash jiddiy muammo bo‘lib turibdi. Oliy ta’lim muassasalari tomonidan ilmiy ishlarni internetga joylashtirib borish internetdan arzon va oson foydalangan holda bunday ishlarni topib olish bu boradagi muammolarni avj olishiga sabab bo‘lmoqda. Albatta xorijda ilmiy matnlarning originalligini tekshiruvchi ko‘plab tijoriy dasturlar borligini yuqorida ham ta’kidlagan edik. Biroq bunga qarshi undanda ko‘proq tijoriy takliflar ham mavjudligi, xususan, “aqlli sinonimizatorlar” hamda dasturni aldash bilan bog‘liq usullarning tinimsiz ishlab chiqarilishi ham jiddiy va dolzarb muammo. Internetdagi onlayn antiplagiat dasturlardan foydalangan holda o‘zbek tilidagi ilmiy matnlarda sodda darak gaplarning mazmuniy o‘xshashlik darajasini aniqlovchi plagiatga qarshi dastur uchun lingvistik baza tayyorlash hamda uning ishlash metodlarini belgilab olish bo‘yicha o‘zimizning taklif va tavsiyalarimizni ishlab chiqdik.

Kompyuter lingvistikasida lingvistik modul termini bugungi kunda muhim ahamiyat kasb etmoqda. Boisi tabiiy tilni kompyuter tiliga o‘tkazilishi, ya’ni kompyuter tizimi vositasida matnga ishlov berish yo‘llarini kashf etish kuzatilmoqda. Bu borada chet tillarining lingvistik dasturlari ishlab chiqilgan va bugungi kunda ular takomillashtirilmoqda. Lingvistik modul ana shunday lingvistik dasturlarning mustaqil tarkibiy qismlari, ya’ni dasturiy ta’minotning muayyan lingvistik jarayonini qamragan qismi hisoblanadi[2]. Aslida til nazariyasi ham mavjud tilning o‘z xususiyatidan kelib chiqqan, uning o‘ziga xos tomonlarini muayyan tartibga ko‘ra tizimli ravishda foydalanish uchun tayyorlab ishlov berishdir. Ya’ni amaliyotdan nazariya kelib chiqqan, bugungi kunda nazariyadan amaliyotga mukammallashib qaytish hodisasi ancha faollashgan.

Matnlarda o‘xshashlikni aniqlash bo‘yicha ***Jakkard o‘lchovi*** (floristik umumiylik koeffitsiyenti, fransuz koeffitsiyenti kommunikatsiya, nemis

olimi)[3]dan foydalaniladi (1901-yilda Pol Jakkard tomonidan taklif qilingan ikkilik o‘xshashlik o‘lchovidir).[1] : KJ = ca + b - c {\displaystyle K_ {J} = {\ frac {c} {a + bc}}} K_ {J} = {\ frac {c} {a + bc}}, bu yerda a – birinchi sinov maydonidagi turlarning soni, b – ikkinchi sinov maydonidagi turlarning soni, c – 1 va 2-saytlar uchun keng tarqalgan turlarning soni. Bu ma’lum bo‘lgan birinchi o‘xshashlik darajasi. Adabiyotda koeffitsiyent muallifining familiyasi, shuningdek Jakkard deb berilgan. Jakkard koeffitsiyenti turli xil modifikatsiyalar va yozuvlardagi ekologiya, geobotanika, molekulyar biologiya, bioinformatika, genomika, proteinomika, informatika va boshqa sohalarda faol qo‘llaniladi. Jakkard o‘lchovi Sorensen o‘lchoviga va cheklangan to‘plamlar uchun Sokal-Snit o‘lchoviga teng.

O‘zbek tilida bitta jumlani turli xil usulda, masalan so‘zlarning joyini o‘zgartirib yoki so‘zlarni sinonimlar bilan almashtirib ifodalash mumkin. Ikki gapning o‘xshashligini aniqlash zarurati kichik amaliy masalani hal qilishda paydo bo‘lgan. Koeffitsiyentlarni aniqlash uchun oddiy o‘lchovlar va yig‘ilgan statistikadan foydalaniladi. Qisqacha aytganda, vazifani quyidagicha ifodalash mumkin: “Yangi jumlalar turli manbalarda ma’lum chastotada keladi. Bir xil fakt bo‘yicha ikkita jumla bo‘lmasligi uchun chiqish ma’lumotlarini filtrlash kerak.”

Ikkita jumlani taqqoslash

Ikkita qatorning o‘xshashlik darajasini aniqlash muammosini hal qilishning bir necha usullari mavjud.

Levenshteyn masofasi[4]

Bir satrni boshqasiga aylantirish uchun qancha operatsiyalarni (qo‘shish, o‘chirish yoki almashtirish) bajarish kerakligini ko‘rsatuvchi raqamni qaytaradi.

Xususiyatlari:

- oddiy dastur;
- so‘z tartibiga bog‘liq;
- chiquvchi natija - bu raqam;
- chiquvchi natijani biror narsa bilan solishtirish kerak.

Masalan, “*Men maktabda dars beraman.”* va *“Men universitetda dars beraman.”* Bu gaplarni bir-biriga moslashishi uchun qancha ko‘p o‘zgartirish qilsa foizi past, qancha kam o‘zgarish bo‘lsa foizi baland chiqadi. Ya’ni *“men”* o‘zgarmadi, *“universitet”* o‘zgardi, *“dars”* o‘zgarmadi, *“beraman”* o‘zgarmadi. Bunda 75 foiz ko‘chirilgan deb beradi. Vaholanki, misol qilingan ikki jumla ma’no-mazmuni jihatidan bir-biridan keskin (universitetdagi o‘qituvchining va universitetdagi ta’limning maktabdan, maktab o‘qituvchisidan ko‘plab farq qiladigan jihatlari mavjud) farq qiladi. “*Men maktabda dars beraman.”* va “*Men maktabda dars beraman”* desa umuman ikki jumlani bir-biriga moslash operatsiyasi amalga oshirilmaydi va o‘z-o‘zidan 100 foiz ko‘chirilgan deb qo‘ya qoladi.

Shingle algoritmi[4]

U matnlarni shingillalarga (inglizcha - tarozilarga), ya’ni 10 so‘zdan iborat zanjirlarga (kesishmalar bilan) ajratadi, shingillalarga xesh funksiyalarni qo‘llaydi, matritsalarni oladi, ularni bir-biri bilan taqqoslaydi.

Xususiyatlari:

- algoritmni amalga oshirish uchun matematik qismini batafsil o‘rganish kerak;
- katta matnlar ustida ishlaydi;
- jumlalar tartibiga bog‘liq emas.

Bu algoritmda, satrlar, so‘zlar solishtirib chiqiladi, so‘z tartibining o‘zgarishi ahamiyatga ega emas, ortiqcha so‘zlarni soniga qarab foizni chiqarib beradi. “*Men maktabda ustozman*.” va *“Men maktabda o‘qituvchiman.”* Bu o‘rinda “ustoz” va “o‘qituvchi”ning sinonim ekanligini bilmaydi va 70 foiz o‘xshash deb chiqaradi.

Jakkard algoritmi

Bu algoritmda esa so‘z emas, harflar taqqoslanadi. *“Men maktabda o‘qituvchi.” va* “*Men maktab o‘qituvchisi.*” Bu o‘rinda –da va –si qo‘shimchalari ortiqcha, 76 foiz o‘xshash deb beradi. Bunga yana misol sifatida asr va asir so‘zlarini 90 foiz o‘xshash deb chiqarishini ta’kidlash orqali jiddiy kamchiliklar bor ekanligini e’tirof etamiz.

Mantiqiy vektorlarning tarkibiy qismlari, ya’ni faqat ikkita 0 va 1 qiymatlarini oladigan komponentlardan foydalanilganda, o‘lchov Tanimoto koeffitsiyenti yoki kengaytirilgan Jakkard koeffitsiyenti sifatida tanilgan. Agar obyektlar turlarning paydo bo‘lishi bilan taqqoslansa (ehtimollik talqini), ya’ni uchrashish ehtimoli hisobga olinadigan bo‘lsa, u holda Jakkard o‘lchovining analogi *Iversen ehtimollik* o‘lchovi bo‘ladi[4].

Bu usulda moslikningina tekshirish imkoniyati mavjud, ya’ni so‘zlar sinonimga o‘zgartirilganda, gap bilan sintaktik aloqaga kirishmaydigan bo‘laklar qo‘shilganda mazmuniy o‘xshashlikni aniqlashda murakkabliklar yuzaga keladi. Ya’ni, bu usulda faqatgina aynan o‘xshash so‘zlargina taqqoslanadi. Bizga esa, aynan o‘xshash bo‘lmagan so‘zlarning semantik o‘xshashligini aniqlash muhim hisoblanadi.

Xulosa o‘rnida shuni aytish mumkinki, ilmiy matnlarning o‘xshashlik darajasini aniqlovchi plagiatga qarshi dasturlar uchun gap bilan sintaktik aloqaga kirishmaydigan bo‘laklarning maxsus ro‘yxatini shakllantirish lozim. Ilmiy matnlardagi sodda darak gaplarning mazmuniy o‘xshashlik darajasini aniqlashda so‘z birikmalarida so‘z va so‘z birikmasi sinonimligi holatini modellashtirishda o‘zbek tilining izohli va sinonimlar lug‘atlaridan foydalanish maqsadga muvofiq[5].

Foydalanilgan adabiyotlar:

1. Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Fil.fan. bo‘yicha falsafa dokt. (PhD) diss. avtoref. – Qarshi, 2018.

2. Elov B., Hamroyeva Sh., Xusainova Z. NLP (tabiiy tilga ishlov berish)ning vazifalari vazamonaviy yondashuvlar. TerDU, FILOLOGIK TADQIQOTLAR: TIL, ADABIYOT, TA’LIM. 2022, 5-6.

1. https://studopedia.info/2-73325.html
2. https://habr.com/ru/post/341148/
3. Murtazayev A. Ilmiy matndagi sodda darak gaplarning mazmuniy o‘xshashligi va lingvistik modellari. – Qo‘qon – 2021. – 133b.