

O‘ZBEK TILI UCHUN SI ASOSIDAGI TAHLIL TIZIMLARINI YARATISHDAGI TILGA XOS MURAKKABLIK

<https://doi.org/10.70728/phil.v05.i06.027>

Ibragimov Javlon Yormatovich
Axborot texnologiyalari va menejment universiteti

Annotatsiya: Mazkur maqolada o‘zbek tili uchun sun’iy intellektga (SI) asoslangan lingvistik tizimlarni yaratish jarayonida uchraydigan asosiy muammolar va ularning lingvistik hamda texnologik sabablari tahlil qilinadi. Tadqiqotda o‘zbek tilining agglutinativ tuzilishi, talaffuz va yozuv o‘rtasidagi nomutanosiblik, kirill va lotin yozuvlarining parallel qo‘llanilishi, erkin so‘z tartibi, pragmatik birliklar, shevalar va frazeologizmlarning sun’iy intellekt tizimlari uchun yuzaga keltiradigan murakkabliklari yoritilgan.

Kalit so‘zlar: sun’iy intellekt, o‘zbek tili, raqamli tilshunoslik, kompyuter lingvistikasi, NLP, avtomatik tarjima, Speech-to-Text (STT), Text-to-Speech (TTS), chatbot, morfologik tahlil

LINGUISTIC CHALLENGES IN CREATING ARTIFICIAL INTELLIGENCE-BASED ANALYSIS SYSTEMS FOR THE UZBEKISTAN LANGUAGE

Javlon Yormatovich Ibragimov
University of information technology and management

Abstract: This article analyzes the main challenges encountered in developing artificial intelligence-based linguistic systems for the Uzbek language, and their linguistic and technological causes. The study highlights the challenges posed to artificial intelligence systems by the agglutinative structure of the Uzbek language, the imbalance between pronunciation and writing, the parallel use of the Cyrillic and Latin alphabets, free word order, pragmatic units, dialects, and phraseological units.

Keywords: artificial intelligence, Uzbek language, digital linguistics, computational linguistics, NLP, automatic translation, speech-to-text (STT), text-to-speech (TTS), chatbot, morphological analysis

Kirish (Introduction). O‘zbek tili uchun SI ga asoslangan tahlil tizimlarini yaratish raqamli tilshunoslikning eng murakkab, ammo eng strategik yo‘nalishidir. SI tilni faqat raqamli ma’lumot emas, tizim, struktura, mantiq, semantika va kontekst sifatida

o'rganadi. Shuning uchun o'zbek tili bo'yicha avtomatik tarjima, STT (Speech-to-Text), TTS (Text-to-Speech), chatbot, grammatik tahlil, semantik tahlil, avtokorreksiya, matn mazmunini ajratish kabi SI vositalarini yaratishda bir qator lingvistik murakkabliklar mavjud. Bu murakkabliklar avvalgi bo'limlarda til infratuzilmasi yoki lug'at tuzilmasi bilan bog'liq muammolardan farq qiladi. Bu yerda muammo tilning o'zida, strukturaviy tabiatida.

Birinchi murakkablik - o'zbek tilining agglyutinativ tuzilishi. Agglyutinativ tillarda ma'no qo'shimchalar orqali yaratiladi, har bir grammatik ma'no, odatda, alohida morfema bilan ifodalanadi. Bu jarayonni SI ga o'rgatish oson emas, chunki kompyuter uchun har bir qo'shimcha alohida semantik yuk ko'taradi. Masalan, "borolmayapsizlarmi" so'zida bir necha grammatik funksiya ketma-ket yuklangan: fe'l o'zagi + harakat majburiyligi + qodirlik yo'qligi + zamon + shaxs-son + so'roq. Bitta SI modeli bunday so'zning morfologik segmentatsiyasini to'g'ri tahlil qilishi uchun tilning qoidalar tizimi emas, qoida ustuvorligini ham bilishi kerak. Ingliz tilida bu muammo deyarli yo'q, chunki grammatik ma'no, asosan, yordamchi so'zlar orqali ifodalanadi ("can not go"), o'zbek tilida grammatik ma'no morfologik darajada birlashadi.

Mavzuga oid adabiyotlarning tahlili (Literature review). Ikkinchi murakkablik - talaffuz va yozuv nomutanosibligi. O'zbek tilida og'zaki nutq bilan yozma shakl o'rtasida sezilarli farq mavjud: talaffuzda qisqarish, urg'u siljishi, ortiqcha tovushlarning tushib qolishi keng uchraydi. Masalan, "borayapman" shakli nutqda "boryapman", "boromman", "boropman", "boryappan" ko'rinishida talaffuz qilinadi. STT (nutqni matnga aylantirish) tizimi uchun bu katta muammo, chunki u matnni to'g'rilangan adabiy shaklda qaytarishi kerak. Agar SI og'zaki variantni to'g'ridan-to'g'ri matnga o'tkazsa, natija noto'g'ri bo'ladi; agar faqat adabiy variantni qabul qilsa, dialektal nutqni "tanish" mexanizmi ishlamaydi. Demak, SI modeli o'zbek nutqidagi fonetik qisqarishlar va talaffuz variantlarini tanish va ularni adabiy shaklga transkodlash qobiliyatiga ega bo'lishi kerak.

Uchinchi murakkablik - kirill va lotin alfavitlarining parallel mavjudligi, ya'ni ikki yozuvli til holati. Bu vaziyat SI uchun qo'shimcha murakkablik tug'diradi: korpuslar bo'linadi, ma'lumotlar ikki tizimga ajraladi, mashina bir xil so'zni ikki xil ko'rinishda ko'rishi mumkin. Masalan, "yo'l" va "йўл", "kitobxon" va "китобхон". Agar SI modeliga ma'lumotlar to'plashda bu ikki yozuv tartibga solinmasa, model so'z chastotalarini noto'g'ri baholaydi, natijada semantik bog'lanishlar buziladi. Mashina o'qitishda ma'lumotlarning yaxlitligi - eng muhim shart, o'zbek tilida ikki yozuv yaxlitlikni buzadi.

To'rtinchi murakkablik - bir so'zning kontekstdan kelib chiqib bir necha so'z turkumiga o'tishi. O'zbek tilida so'zlar juda ko'p holatda "ko'chib yuradi". Masalan, "yashil" sifat, ammo "yashillar" - ot, "yashillashmoq" - fe'l. Ingliz tilida ham bunday

o'tishlar mavjud, lekin kontekstga bog'liq aniqlik ko'p hollarda sintaksis orqali beriladi. O'zbek tilida bunday o'tishlar morfologik yondashuv va kontekstual semantika bilan belgilanadi. SI uchun bu juda murakkab: model so'zning "yashil = rang belgilovchi sifat" ekanligini bilishi mumkin, ammo "yashillar = siyosiy guruh atamasi" ma'nosini semantik maydon asosida aniqlashga to'g'ri keladi.

Tadqiqot metodologiyasi (Research Methodology). Beshinchi murakkablik - so'z tartibining erkinligi. O'zbek tilida SOV tartibi (ega + to'ldiruvchi + kesim) asosiy, lekin stilistik yoki pragmatik maqsadga ko'ra tartib o'zgarishi mumkin:

- "Men kitobni o'qidim".
- "Kitobni men o'qidim".
- "O'qidim men kitobni".

SI uchun bunday o'zgaruvchanlik g'ayrioddiy, chunki ko'pgina NLP modellari qat'iy tartibli tillarda o'qitilgan (SVO - ingliz tili). Demak, SI o'zbek tilida gap tarkibini belgilash uchun sintaksisga emas, morfologik ko'rsatkichlarga tayanishi kerak bo'ladi.

Oltinchi murakkablik - o'zbek tilidagi modal ma'nolar va so'z-gaplar. "Bo'ldi", "xo'p", " ", "deylik" kabi birliklar grammatik jihatdan gapga o'xshamaydi, lekin pragmatik vazifani bajaradi. SI uchun bularning intonatsiyasi, kontekstdagi funksiyasi, so'z-gap bo'lishi - formal qoidalarga sig'maydi. Chatbotlar o'zbek tilida aynan shuning uchun tabiiy javob qaytara olmaydi: ular pragmatikani tushunmaydi.

Yettinchi murakkablik - so'z yasaliishi va qo'shimchalar orqali konnotativ ma'no ifodalanishi. Masalan, "-vchan", "-garchi", "-day", "-ona" kabi qo'shimchalar konnotativ-estetik ma'no hosil qiladi:

- bolalarcha quvonch
- do'stona yondashuv

Bu ma'nolar lug'atga tushmaydi, SI estetik semantikani faqat statistik model orqali o'rganishi mumkin.

Bu bo'limda ta'kidlanishi kerak bo'lgan eng muhim fenomenlardan biri - shevalarning mavjudligi. O'zbek tili ko'plab dialekt va shevalarga ega: Farg'ona, Qashqadaryo, Surxondaryo, Buxoro, Xorazm, Qo'qon, Qarshi, Toshkent, Samarqand... Har bir sheva fonetik, leksik va morfologik jihatdan farq qiladi. Nutqni matnga aylantirish tizimlari uchun bu juda katta muammo:

"borayapti" → "boryapti" → "borvotti".

SI qaysi variantni standart deb qabul qilishi kerak? Shevalarni tanimaydigan STT tizimi noto'g'ri transkripsiya qiladi, keyin TTS noto'g'ri talaffuz qiladi. Bu matndan nutqqa qayta kodlashning uzluksizligini buzadi.

Semantik modellarni yaratishda yana bir to'siq - frazeologizmlar. O'zbek tili frazeologizmlarga boy: "tosh qotgan", "u hammani yeb qo'yadi", "boshi osmonga yetdi". SI tizimi bularni so'zma-so'z tarjima qiladi, semantik ma'nosini tushunmaydi.

Bu muammoni hal qilish uchun frazeologik birliklar semantik ontologiya bilan belgilanishi kerak - hozircha bunday ontologiya yo'q.

SI tizimlarini yaratishda eng katta to'siqlardan biri - SI o'qitilishi uchun kerak bo'lgan ma'lumotlar hajmining yetishmasligi. Chet tillari uchun milliardlab so'z mavjud bo'lsa, o'zbek tilida sifatli, teglangan, standartlashtirilgan korpus juda cheklangan. SI qancha ko'p matn ko'rsa, shuncha yaxshi o'rganadi; o'zbek SI ma'lumot yetishmasligi sababli o'rgana olmaydi. Bu jarayonni murakkablashtiruvchi yana bir omil - ta'lim resurslarining yopiq bo'lishi. Raqamli platformalarning ko'pchiligi grant loyihalari sifatida yaratiladi va keyin yopiladi. SAgamosh o'quv ma'lumotlari jamlanmaydi, u ekspertlar qo'lida tarqoq holda qoladi. Oxir-oqibat SI modeli yaratilgan taqdirda ham, u o'zbek tili foydalanuvchilarining madaniy konteksti va nutq odatlarini anglashga qodir bo'lmasligi mumkin. Masalan, "bormoqchimisiz yoki kelaymi?" jumlasida taklif, hurmat, muomala bor. SI buni buyruq deb tushunishi mumkin. SI o'zbek tilida muvaffaqiyatli ishlashi uchun tilni faqat algoritm sifatida emas, tirik ijtimoiy muhit sifatida ko'rish zarur. O'zbekcha SI yaratish texnologik emas, tafakkur masalasidir.

Tahlil va natijalar (Analysis and results). Raqamli davr filologiyani yangicha sinov oldiga qo'ydi: endi tilni o'rganish faqat matn tahlili yoki grammatikani tushuntirish emas, uni texnologik muhitda qayta qurish ham demakdir. Shu jarayonda filologik tafakkur va texnologik tafakkur o'rtasida keskin tafovut paydo bo'ldi. Bir tomonda so'z, ma'no, kontekst bilan ishlaydigan insonparvar tafakkur; ikkinchi tomonda algoritm, kod, ma'lumot strukturalarini ko'radigan texnik tafakkur. Aslida tilning kelajagi bu ikki yo'nalishning qarama-qarshiligida emas, ularning uyg'unligida hal bo'ladi. Chunki raqamlashtirish texnologik jarayon bo'lsa-da, u filologik mazmunni saqlab qolmasa, insoniy ma'noni yo'qotadi. Demak, bugungi asosiy muammo texnologiyani biladigan, lekin ma'noni his etadigan mutaxassislar avlodini yetishtirishdir. Zero, "Raqamlashtirilgan tilda inson tilining haqiqiy nafasi va nozikligini yo'qotmasligi muhimdir"[1].

Filologik tafakkur tilni inson tajribasining in'ikosi sifatida ko'radigan, so'z orqali tafakkur, madaniyat va tarixni tahlil qiladigan idrok shaklidir. Texnologik tafakkur ma'lumotni kod, tizim, model sifatida qayta ishlaydi. Birinchisi ma'noni yaratadi, ikkinchisi uni qayta ishlab chiqaradi. Shuning uchun ularning birlashuvi tilning raqamli barqarorligini ta'minlovchi yagona yo'ldir. Hozir o'zbek tilshunosligida bu integratsiya jarayoni endigina bosqichma-bosqich shakllanmoqda. Universitet, ilmiy markaz va AT-loyihalarda hali filologik bilim va texnik kompetensiya orasidagi farq juda chuqur. Tilshunoslar texnologik jarayonni "begona mexanizm" sifatida ko'radi, dasturchilar tilni "ma'lumot bazasi" darajasida idrok etadi. Shu ziddiyat tufayli raqamli til loyihalari ko'pincha na ilmiy chuqurlikka, na texnik mukammallikka erishadi.

Filologik tafakkurni texnologiya bilan bog'lash uchun, avvalo, ta'lim paradigmasi o'zgarishi zarur. Filologlar endi dasturlash asoslarini, ma'lumotlar tuzilmasini, korpus lingvistikasi metodlarini bilmasdan raqamli loyihada ishtirok eta olmaydi. Zamonaviy dunyoda muloqot va til raqamlashishga asoslanadi[2]. Xuddi shunday, AT mutaxassislari ham tilning sintaktik, semantik, pragmatik darajalarini tushunmasdan tilga oid mahsulot yarata olmaydi. Demak, integratsiya o'quv jarayonidan boshlanishi kerak: filologlar uchun "Raqamli tilshunoslik", "Til muhandisligi", "Lingvistik ma'lumotlar bazasi", "Korpus texnologiyalari" kabi yangi fanlar, texnik yo'nalishlarda "Natural Language Processing asoslari", "Kompyuter lingvistikasiga kirish" kabi darslar joriy qilinishi zarur. Bu ikki tafakkurning uchrashuvi nafaqat o'quv rejalarda, fikrlash madaniyatida ham sodir bo'lishi kerak.

Integratsiyaning asosiy to'siqlaridan biri metodologik qarama-qarshilikdir. Filologik tafakkur individual tahlil, kontekst, his va ma'no izlanishiga tayanadi. Texnologik tafakkur umumlashtirish, standartlashtirish, avtomatlashtirishni talab qiladi. Tilning tabiati aynan bu ikki yondashuvni birlashtirishni talab etadi. Masalan, bir so'zning ma'nosi kontekstdan kelib chiqadi ("ko'r" - fe'l, sifat, buyruq shakli), buni SI faqat statistik modellar orqali bilib ola oladi. Shuning uchun filolog ma'noni tushuntiradi, dasturchi uni modelga joylaydi. Bu ikki tafakkur bir-birini to'ldiradi: filolog kontekstni beradi, texnolog algoritmik shaklni. Bu birlashuv orqali til haqidagi bilim "inson tafakkuri"dan "raqamli ong"ga o'tadi.

Integratsiyaning yana bir muhim yo'nalishi "til muhandisi" kasbini shakllantirishdir. Til muhandisi - filologik bilimni texnologik tizimlarga tatbiq qila oladigan mutaxassis. U til strukturasi, semantik tarmoq, matn tahlili va algoritmik modellashtirishni bir butun jarayon sifatida ko'radi. Bu kasb G'arbda "Computational Linguist" yoki "Language Engineer" sifatida allaqachon mustahkam o'rin egallagan. O'zbekistonda bu soha endigina shakllanmoqda. Til muhandisi o'zbek tili uchun TTS/STT tizimlari, korpuslar, SI-tarjimon, avtomatik tahlilchi va matn filtrlari yaratishda kalit rol o'ynaydi. Ammo bu mutaxassisni tayyorlash uchun hozircha akademik asoslar yetarli emas: filologik fakultetlarda algoritmik tafakkur o'rgatilmaydi, texnik fakultetlarda lingvistik bilimlar o'quv dasturidan chetda qolgan.

Xulosa va takliflar (Conclusion/Recommendations). O'zbek tilini raqamlashtirish bugungi ilmiy-texnologik rivojning ajralmas qismiga aylandi. Tilni faqat aloqa vositasi sifatida emas, raqamli tafakkur tizimining markaziy elementi sifatida tushunish davri keldi. Yutuqlar bor - elektron lug'at, dastlabgi korpus, SI tahlilchilar; ammo bu jarayon hali tizimga solinmagan, muvofiqlik, metodik aniqlik va integratsiya yetishmaydi. Asosiy vazifa tilni texnologiyalash emas, texnologiyani tilga bo'ysundirish. Buning uchun filologik tafakkur va texnologik tafakkur bir yo'nalishda ishlashi, yagona milliy raqamli platforma yaratilishi, har bir resurs ochiq, standartlashtirilgan va o'zaro bog'liq bo'lishi zarur. O'zbek tili SI tizimlarida mustaqil "fikrlay" oladigan darajaga chiqishi

uchun lingvistik asos, metodika va intellektual kuch birlashmog'i kerak. Raqamlashtirish jarayoni texnik muammo emas, tafakkur yangilanishidir. Shuning uchun har bir filolog raqamli jarayonda o'z mas'uliyatini his qilishi, o'zbek tilini kelajak texnologiyalarining markaziga aylantirishga intilishi lozim.

Foydalanilgan adabiyotlar

1. Stanciu N., Jager S.D. Digital language // A Multilingual Language in It's Multiple Dimensions. 2024. - 284 p.
2. Burada M., Tatu O., Sinu R. Language and Communication in the Digital Age. - Cambridge: Cambridge Scholars Publishing, 2023. - 378 p.
3. Burada M., Tatu O., Sinu R. Language and Communication in the Digital Age. - Cambridge: Cambridge Scholars Publishing, 2023. - 378 p.
4. Fuertes-Olivera., Bergenholtz H. e-Lexicography. - London: A&C Black, 2011. - 341 p.