UDK: 81.33

KAM RESURSLI TILLARDA G2P MODEL TAYYORLASHDA FONETIK KORPUSLARNING AHAMIYATI: O‘ZBEK TILI MISOLIDA

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Maxmudjonova Gulshaxnoz Ulug‘bek qizi,
tayanch doktorant
gulshaxnozmahmudjonova@gmail.com
ToshDO‘TAU

Annotatsiya: Ushbu maqolada kam resursli tillar, xususan o‘zbek tili uchun Grafema-fonema (G2P) modelini yaratishda fonetik korpuslarning o‘rni va ahamiyati yoritilgan. G2P konversiyasi – matnni fonemalarga aylantirish jarayoni bo‘lib, zamonaviy nutq sintezi (TTS) tizimlarining asosiy komponentlaridan biridir. O‘zbek tili fonetik tizimining murakkabligi, undagi tarixiy, fonologik va morfonologik xususiyatlar, shuningdek, dialektal tafovutlar fonetik korpuslar asosida formallashtirilmaguncha, modelning sifatli ishlab chiqilishi mushkul bo‘ladi.

Abstract: This article explores the significance and role of phonetic corpora in developing Grapheme-to-Phoneme (G2P) models for low-resource languages, with a particular focus on the Uzbek language. G2P conversion – the process of mapping text to phonemes – is a core component of modern Text-to-Speech (TTS) systems. Due to the complexity of Uzbek phonetics, including its historical, phonological, and morphonological features, as well as dialectal variation, a high-quality G2P model cannot be effectively developed without the formalization of phonetic corpora.

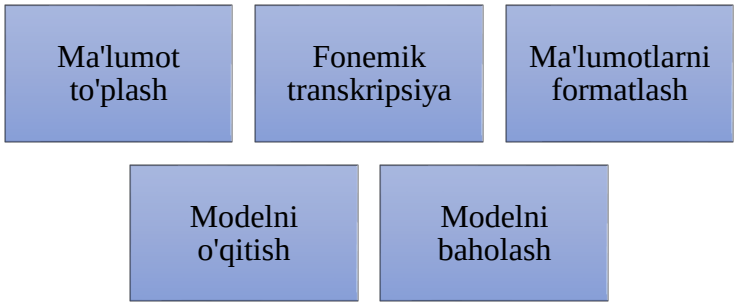
Аннотация: В данной статье рассматриваются значение и роль фонетических корпусов при создании моделей преобразования графем в фонемы (G2P) для языков с ограниченными ресурсами, в частности, узбекского языка. Преобразование G2P – это процесс трансформации текста в фонемы, который является ключевым компонентом современных систем синтеза речи (TTS). Сложность фонетической системы узбекского языка, включая её исторические, фонологические и морфофонологические особенности, а также диалектные различия, делает невозможным создание качественной G2P-модели без формализации фонетического корпуса.

Kalit so‘zlar: *fonetik korpus, grafema, transkripsiya, alifbo, lug‘at, ma’lumotlar bazasi*

G2P (grafemadan fonemaga o‘tkazish) modeli matndagi so‘zlarni ularning talaffuziga avtomatik konvertatsiya qilish vazifasini bajaradi. Bu texnologiya nutqni sintezlash (Text-to-Speech, TTS) va nutqni avtomatik tanish (Automatic Speech Recognition, ASR) tizimlarida muhim rol o‘ynaydi[1]. Masalan, TTS jarayonida tovushlarning to‘g‘ri chiqarilishini ta’minlash uchun matnni fonetik transkripsiyaga aylantirish zarur bo‘ladi – ayniqsa yangi yoki lug‘atda yo‘q so‘zlarda (Out-of-vocabulary, OOV)[2]. Yetakchi xalqaro tillar (ingliz, fransuz, rus) uchun bunday modellar mavjud bo‘lib, ularni tayyorlashda keng qamrovli talaffuz lug‘atlari va fonetik korpuslardan foydalaniladi[3]. Biroq, kam resursli tillarda – jumladan, o‘zbek tilida – bunday fonetik ma’lumotlar bazasi yo’qligi G2P modelini yaratishni qiyinlashtiradi.

O‘zbek tili dunyoda o‘n million insonlar tomonidan so‘zlashiladi (turli baholashlarga ko‘ra 30-40 million atrofida[4] va turkiy tillar orasida ikkinchi eng ko‘p so‘zlovchiga ega tildir. Shunga qaramay, hozirgi paytda o‘zbek tilida talaffuz lug‘atlari yoki fonetik belgilangan matn korpuslari deyarli mavjud emas. Natijada, G2P modelining aniqligi uchun zarur bo‘lgan orfografiya – tovush mosliklarini mashinaga o‘rgatish murakkablashadi. Masalan, bir necha tillar uchun yaratilgan ko‘p tilli G2P tizimlari[5] o‘zbek tilini ham qamrab olgan bo‘lsa-da, ularning resurslari to‘liq ishonchli emas: CharsiuG2P loyihasi o‘zbek tili bo‘yicha ishlatilgan talaffuz lug‘atida xatoliklar borligini va uni yangilash zarurligini qayd etgan[6]. Bu holat kam resursli tilga xos bo‘lgan muammo – *fonetik ma’lumotlarning sifati va standartlashmagani* – modelga bevosita ta’sir qilishini ko‘rsatadi.

Fonetik korpus tuzish bosqichlari (1-rasmga qarang). O‘zbek tili uchun G2P modelini yaratishda birinchi navbatda fonetik korpus – ya’ni so‘zlar va ularning to‘g‘ri fonemik transkripsiyasidan iborat ma’lumotlar to‘plami – shakllantiriladi. Korpus tuzish jarayoni quyidagi bosqichlarni o‘z ichiga oladi:



1-rasm. Fonetik korpus tuzish bosqichlari

Ma’lumot to‘plash: avvalo, o‘zbek tilida reprezentativ matnlar to‘plami yig‘iladi. Bu jarayonda adabiy va og‘zaki nutqni qamrab oluvchi turli janr va uslubdagi matnlar (rasmiy hujjatlar, yangiliklar, badiiy asarlar, og‘zaki nutq namunalari) tanlanadi. Maqsad – korpusda o‘zbek tilidagi barcha asosiy tovush

birikmalari va holatlarini qamrab olishdir. Korpusga kiritiladigan so‘zlar soni imkon qadar ko‘proq va xilma-xil bo‘lishi lozim (masalan, 50 – 100 ming atrofida so‘z shakllari). Yig‘ilgan xom matnlar tozalash va normalizatsiya qilinadi: keraksiz belgi va shum (masalan, raqamlar, qisqartmalar) chiqarib tashlanadi, yoki standart ko‘rinishga keltiriladi. So‘zlar alohida leksik birliklar sifatida ajratiladi va korpusning leksik ro‘yxati shakllantiriladi.

Fonemik transkripsiya va annotatsiya: keyingi bosqichda korpusdagi har bir so‘zning fonemik transkripsiyasi tayyorlanadi. Buning uchun transkripsiya standarti tanlanishi zarur. Ilmiy amaliyotda xalqaro IPA (International Phonetic Alphabet) keng qo‘llaniladi, chunki u barcha tovushlarni unifikatsiyalangan belgilarda ifodalash imkonini beradi. O‘zbek fonetikasi uchun moslashtirilgan alifbo ham qo‘llaniladi. Transkripsiyani lingvist-mutaxassislar qo‘lda bajarishi yoki maxsus yarim-avtomatik vositalar yordamida amalga oshirishi mumkin. Hozirgi kunda mavjud ba’zi dasturlar (Epitran, CharsiuG2P, MFA) boshqa tillardagi modellarga asoslanib o‘zbekcha taxminiy transkripsiya berishi mumkin, lekin ularni ehtiyotkorlik bilan qo‘llash lozim – yuqorida ta’kidlanganidek, ko‘p tilli model bergan transkripsiya xatolarini inson tomonidan tuzatish talab etiladi. Demak, annotatsiya jarayoni sifat jihatdan nazoratlangan bo‘lishi kerak: har bir so‘zning transkripsiyasi kamida ikki mutaxassis tomonidan tekshirilishi, bir xil transkripsiya me’yoriga rioya etilishi zarur. Annotatsiyada fonemalar izchil belgilanishi (bir xil tovush har doim bir xil belgi bilan ifodalanishi shart), urg‘u yoki intonatsiya kabi qo‘shimcha prosodik belgilar zarur bo‘lsa, alohida belgi bilan ko‘rsatilishi mumkin.

Ma’lumotlarni formatlash va bazaga kiritish: transkripsiyalangan korpus ma’lumotlari lug‘at shaklida tartiblanadi, har bir qatorda so‘z va uning fonetik transkripsiyasi yozilgan format (masalan, JSON ko‘rinishda) saqlanadi. Bu lug‘at G2P modelini o‘qitish uchun tayyor ma’lumotlar bazasi vazifasini bajaradi. Korpus tuzishda, shuningdek, morfologik belgilarni ham saqlab qo‘yish foydali bo‘ladi, chunki o‘zbek tili agglyutinativ til bo‘lib, morfemalar chegarasida tovush o‘zgarishlari kuzatilishi mumkin. Korpusning har bir yozuvi ko‘rinishi jihatdan: *“so‘z – [fonemik transkripsiya]”* tarzida bo‘ladi. Misol: kitob — [k i t ɒ b], quyosh — [q u j ɒ ʃ] kabi.

Modelni o‘qitish (tayyorlash): tayyor fonetik korpus asosida G2P modeli o‘qitiladi. G2P modellarini ishlab chiqishda ikki asosiy yondashuv mavjud: lug‘atga asoslangan va qoidaga asoslangan usullar. Lug‘atga asoslangan usulda katta hajmdagi tayyor talaffuz lug‘ati modelga yuklanadi va undan statistik yoki neyron tarmoq usullari bilan yangi so‘zlar uchun talaffuz generatsiya qilinadi. Qoidaga asoslangan yondashuvda esa tilshunoslikdagi fonologik qoidalar dasturlashtiriladi va faqat istisno so‘zlar uchun alohida lug‘at yuritiladi. Hozirgi zamon G2P modellari ushbu yondashuvlarning kombinatsiyasidan foydalanadi. Masalan, Phonetisaurus[7] kabi vositalar WFST asosida lug‘atni o‘rgatib, keyin yangi so‘zlarni statistik

qoidalar yordamida transkripsiya qiladi[8]. NVIDIA NeMo yoki CharsiuG2P kabi platformalar esa neyron tarmoqqa asoslangan yondashuvdan foydalanadi (masalan, Transformer arxitekturasi yoki ByT5 modeli) va bir nechta til uchun birgalikda o‘qitilib, keyin ma’lum bir tilga moslashtiriladi. Bizning holatda, korpus nisbatan kichik va tilimiz xususiyatlari o‘ziga xos bo‘lgani uchun, qoida-lug‘atli model konsepsiyasini qo‘llash maqsadga muvofiq: ya’ni, o‘zbek tilining fonetik qonuniyatlarini inobatga olgan holda, kichik lug‘at va qoidalar integratsiyasi bilan ishlaydigan modelni tanlash. Bu jarayonda Phonetisaurus yoki MFA[9] singari vositalar yordamida avval grafema-fonema moslashtirish (alignment) amalga oshirilib, so‘ng n-gram model yoki neyron tarmoqni o‘qitish orqali grafema ketma-ketliklaridan fonema ketma-ketliklariga xarita hosil qilinadi. Modelni tayyorlashda korpusning 80-90% qismi o‘qitish uchun, qolgan qismi test va tasdiqlash (validation) uchun ajratiladi. Natijada, model yangi berilgan so‘zning imlosidan uning talaffuzini bashorat qilishni o‘rganadi.

Modelni baholash: G2P model sifatini o‘lchash uchun fonemik xato darajasi (Phoneme Error Rate, PER) yoki so‘z bo‘yicha aniqlik kabi mezonlar qo‘llanadi. Tayyor model korpusdan ajratib olingan test to‘plamidagi so‘zlar uchun transkripsiya yaratadi va ular mutaxassis tayyorlagan etalon transkripsiya bilan taqqoslanadi. Kam resursli til sharoitida modelning dastlabki natijalari kutilganidek pastroq bo‘lishi mumkin; bu holda modelni takomillashtirish uchun korpusni kengaytirish (yangi so‘zlar qo‘shish), yoki model arxitekturasini murakkablashtirish kabi choralar ko‘riladi. Shu tariqa, metodologiya bosqichlari takomillashtirib borilishi ham mumkin – masalan, model chiqishlaridagi tizimli xatolar tahlil qilinib, ularni tuzatish uchun korpus annotatsiyasiga qo‘shimcha qoidalar yoki misollar kiritiladi.

Fonemik transkripsiya standarti va annotatsiya prinsiplari. Fonetik korpusni annotatsiya qilishda fonemik transkripsiya standartining to‘g‘ri tanlanishi va izchil qo‘llanishi hal qiluvchi ahamiyatga ega. O‘zbek tilining tovush tizimi uchun alifbo IPA standartiga moslashtirilishi kerak. Transkripsiyada unli fonemalar uchun IPA belgilarini qo‘llaganda, ularning allofonik variantlari ham e’tiborga olinadi. O‘zbek tilida ba’zi kombinator hodisalar yuz beradi, ular transkripsiya me’yorlarida aks ettirilishi lozim. Masalan, jarangli/jarangsizlanish assimilatsiyasi: so‘z ichida yoki qo‘shimcha qo‘shilganda, ba’zan yonma-yon undoshlar bir-biriga ta’sir qiladi. Transkripsiya jarayonida barcha qoidalar ishlab chiqilib, bir xil uslubda bajarilishi kerak. Bu degani, biror tovushni ifodalash uchun qanday belgi tanlansa, hamma joyda shuni qo‘llash lozim; birikma yoki qisqartmalarning talaffuzi bo‘yicha formallashtirilib, qat’iy rioya qilinadi. Masalan, “SMS” so‘zi harflab o‘qiladi (es-em-es) yoki “YUNESKO” qisqartmasi [Yunesko] deb o‘qiladi – buni oldindan hal qilib, korpusga to‘g‘ri kiritish kerak. Annotatsiya prinsiplari bo‘yicha qo‘llanma tuzilib, ularda turli murakkab holatlar uchun yechimlar ko‘rsatiladi. Bu prinsipial

jihatlar model sifatiga katta ta’sir qiladi – agar bir xil so‘z turlicha transkripsiya qilinsa, model buni shovqin sifatida qabul qilib, noto‘g‘ri o‘rganadi.

Fonetik korpus sifati va model natijalari. G2P modelining muvaffaqiyati ko‘p jihatdan tayyorlangan fonetik korpusning sifatiga bog‘liq. Agar korpusda xatolar yoki nomuvofiqliklar bo‘lsa, model noto‘g‘ri tarzda o‘rganadi va bu talaffuz xatolariga olib keladi. Shu bois, korpus annotatsiyasida qo‘yilgan qoidalarga qat’iy amal qilish va uni imkon qadar to‘liq va xatosiz qilish muhim. Misol uchun, o‘zbek tilida “o” va “o‘” fonemalarini adashtirish modelni chalg‘itishi va natijada “ol” va “o‘l” so‘zlarini farqlay olmaslikka olib kelishi mumkin. Shunday holatlarda model chiqishini tekshirib, korpusdagi potentsial xatoni tuzatish lozim.

Annotatsiya muammolarining yana biri – inson omili. Fonetik transkripsiya qo‘lda qilinganida, ayniqsa katta hajmda, ba’zan subyektiv farqlar paydo bo‘lishi mumkin. Turli lingvistlar bir tovushni turlicha talqin qilishlari yoki diakritik belgilar qo‘llashdagi uslubi farq qilishi mumkin.

Fonetik korpus hajmi ham ahamiyatli. Dastlabki bosqichda korpus kichik bo‘lsa, model yetarlicha qoidalarni o‘rganmasligi mumkin. Tadqiqotlar shuni ko‘rsatadiki, Transformer kabi zamonaviy modellar katta ma’lumot talab qiladi va bitta tilning ma’lumotlari kam bo‘lsa, ko‘p tilli o‘qitish orqali samaradorlikni oshirish mumkin. Bizning kontekstimizda ham agar o‘zbek tilida 50 ming so‘zli korpus yetarli natija bermasa, uni turkiy tillar (masalan, qozoq, turk, uyg‘ur) ma’lumotlari bilan birga o‘qitib, keyin o‘zbekchaga moslashtirish (fine-tuning) strategiyasini qo‘llash mumkin. SIGMORPHON kabi tanlovlarda ham kam resursli tillar uchun transfer learning yondashuvi samarali ekani ko‘rsatilgan[10]. Biroq, bunday ko‘p tilli modelda har bir tilning o‘ziga xos fonetikasi “yo‘q bo‘lib ketmasligi” uchun ehtiyot bo‘lish zarur – masalan, turk tilidagi unlilar uyg‘unlashuvi qoidalari modelga singib, o‘zbek tiliga noto‘g‘ri tatbiq bo‘lmasligi kerak. Shu sababdan biz taklif qilayotgan yondashuv fonetik jihatdan asoslanadi: ya’ni, modelga boshidan o‘zbek tilining fonologik xususiyatlari (masalan, unlilar uyg‘unlashuvi yo‘qligi, qator va tor unlilar farqi, g‘/q tovushlari mavjudligi kabi) haqida ma’lumot beriladi yoki qoidalar shaklida kiritiladi. Bu bilimlar modelning ichki qatlamlarida yoki qoidalovchi modulida hisobga olinsa, u holda korpus kichikroq bo‘lsa ham, model ma’lum qoidalarga tayangan holda to‘g‘riroq ishlaydi. Masalan, yapon tilida kam resurs sharoitida mutaxassislar aynan qoidabazali G2P tizimini taklif qilishgan – lingvistik adabiyotga asoslanib, maxsus qoidalar to‘plami tuzilgan va u ko‘p holatlarda neyron modelga teng natija bergan[11]. O‘zbek tilida ham xuddi shunday, tilshunoslikka asoslangan yondashuvni joriy etish mumkin: bu modelga fonetik korpusdan tashqari fonologik qoidalarni qo‘lda kiritishni anglatadi.

Fonetik korpusning roli. Kam resursli til – xususan, o‘zbek tili – uchun G2P modelini muvaffaqiyatli yaratishda fonetik korpus hal qiluvchi ahamiyatga ega.

Talaffuzi aniq belgilangan so‘zlar jamlanmasi modelni to‘g‘ri o‘qitish va tekshirish imkonini beradi. Korpus hajmi va sifati qanchalik yuqori bo‘lsa, model natijalari shunchalik ishonchli chiqadi. Korpus yaratishda barcha asosiy tovush birikmalari va fonologik hodisalarni qamrab olish lozim, aks holda model notanish holatlarda xatoga yo‘l qo‘yishi mumkin[12].

Yakunda, o‘zbek tili misolida ko‘rsatildiki, kam resursli til uchun G2P model yaratish tilshunos va sun’iy intellekt mutaxassislarining yaqin hamkorligini talab etadi. Fonetik korpus shakllantirish va uni tahlil qilish orqali biz modelning birinchi talqinini yaratishimiz mumkin. Keyingi amaliy ishlar bosqichida mazkur nazariy ishlab chiqilgan yondashuv asosida G2P modelining dastlabki amaliy tadbiqi amalga oshiriladi va uning natijalari baholanadi. Ushbu konseptual ish kelgusida nafaqat o‘zbek tili, balki unga o‘xshash ahvoldagi boshqa kam resursli tillar (masalan, tojik, qirg‘iz, turkman) uchun ham foydali bo‘lishi mumkin – zero, til texnologiyalari sohasida resurslarni yaratish va ulardan samarali foydalanish globallashgan dunyoda kichik tillarning rivoji uchun muhim ahamiyat kasb etadi.

Foydalanilgan adabiyotlar:

1. Irie E., Samson J., Saee Suhaila. A Review on Grapheme-to-Phoneme Modelling Techniques to Transcribe Pronunciation Variants for Under-Resourced Language. Pertanika Journal of Science and Technology. 2023.
2. NVIDIA. NeMo Toolkit Documentation. – URL: https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/tts/g2p.html
3. Kim B. C., Lee G. B., Lee J. H. Hybrid Grapheme-to-Phoneme Conversion for Unlimited Vocabulary. Department of Computer Science & Engineering, Pohang University of Science & Technology, Pohang, Korea. 1999
4. en.wikipedia.org
5. CharsiuG2P – URL: https://github.com/lingjzhu/CharsiuG2P
6. El-Hadi, C., Mhania, G. Phonetisaurus-based letter-to-sound transcription for Standard Arabic. In 2017 5th International Conference on Electrical Engineering - Boumerdes. 2017
7. Phonetisaurus – URL: https://github.com/AdolfVonKleist/Phonetisaurus
8. github.com
9. Montreal Forced Aligner – URL: https://github.com/MontrealCorpusTools/mfa-models
10. Salatas, J. Phonetisaurus: A WFST-driven Phoneticizer – Framework Review. Retrieved from https://www.johnsalatas.com/phonetisaurus-wfst-review, 2012
11. Jiampojamarn S. Grapheme-to-Phoneme Conversion and Its Application to Transliteration. – University of Alberta, 2009.
12. Cheng S., Zhu P., Liu J., Wang Z. A Survey of Grapheme-to-Phoneme Conversion Methods // *Applied Sciences*, 2024.