



UDK: 81'33:81'23:81-13

*Nilufar ABDURAXMONOVA,*  
*Filologiya fanlari doktori, professor*  
*O'zbekiston Milliy universiteti, Toshkent, O'zbekiston*  
*E-mail:n.abduraxmonova@nuu.uz*  
**Sohiba ABDURAXMONOVA,**  
*O'zbekiston Milliy universiteti tayanch doktoranti*

*Renessans universiteti professori, f.f.d L.Raupova taqrizi asosida*

### BOLALAR NUTQI KORPUSINI YARATISHNING NAZARIY VA METODOLOGIK ASOSLARI

Annotatsiya

Ushbu maqola o'zbek bolalar nutqi korpusini yaratishning nazariy va metodologik muammolariga bag'ishlangan. Shuningdek, ma'lumotlarni yig'ish jarayoni, ularni raqamli formatga o'tkazish va o'zbek tili grammatikasidan kelib chiqqan holda teglash jarayonlari yoritilgan. Shunga ko'ra dunyo standartlariga javob beruvchi CHILDES corpus tizimidan foydalanish samaradorligi ko'rsatib berilgan.

**Kalit so'zlar:** Bolalar nutqi korpusi, CHILDES, CHAT, MLUm, agglutinativ tillar, morfologik annotatsiya, CLAN, ELAN.

### THEORETICAL AND METHODOLOGICAL BASIS OF CREATING A CHILDREN'S SPEECH CORPUS

Annotation

This article is devoted to the theoretical and methodological problems of creating a corpus of Uzbek children's speech. It also covers the process of collecting data, converting them into a digital format, and tagging them based on the grammar of the Uzbek language. Accordingly, the effectiveness of using the CHILDES corpus system, which meets world standards, is demonstrated.

**Key words:** Children's speech corpus, CHILDES, CHAT, MLUm, agglutinative languages, morphological annotation, CLAN, ELAN.

### ТЕОРЕТИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ ДЕТСКОГО ТЕОРЕТИЧЕСКОГО КОРПУСА

Аннотация

Данная статья посвящена теоретическим и методологическим проблемам создания корпуса речи узбекских детей. В ней также рассматривается процесс сбора данных, их преобразования в цифровой формат и разметки на основе грамматики узбекского языка. Соответственно, демонстрируется эффективность использования корпуса CHILDES, соответствующего мировым стандартам.

**Ключевые слова:** Корпус детской речи, CHILDES, CHAT, MLUm, агглютинативные языки, морфологическая аннотация, CLAN, ELAN.

**Kirish.** Bizga ma'lumki, zamonaviy tilshunoslikda xususan, psixolingvistikada bolalar nutqini o'rganish zaruriy va dolzab masalalardan biri sanaladi. Bolalar nutqining shakllanishi inson ontogenezidagi eng murakkab lingvistik jarayondir. Axborot texnologiyalarining tilshunoslikka kirib kelishi natijasida bolalar nutqini an'anaviy kuzatishdan ko'ra, statistik ma'lumotlarga asoslangan korpus lingvistikasi metodlariga ehtiyoj ortmoqda. Bunga quyidagi sabablarni ko'rsatishimiz mumkin:

Inson xotirasi va oddiy yozuvlar bolaning kun davomida aytadigan minglab gaplarini qamrab ololmaydi. Korpus esa o'nlab, yuzlab bolalarning yillar davomida aytgan millionlab so'zlarini saqlash imkonini beradi.

Eksperimental laboratoriya sharoitida bola o'zini erkin tutmasligi mumkin. Korpus uchun materiallar bolaning uyida, o'yin davomida, tabiiy muloqotda olinadi.

Ingliz va o'zbek bolasi nutqini solishtirish uchun ikkala tilning ham raqamli bazasi (korpusi) bo'lishi shart. Bu esa tillardagi nutq o'zlashtirish jarayonini qiyosiy solishtirish imkonini beradi.

**Mavzuga oid adabiyotlar tahlili.** Bugungi yondashuvlarga ko'ra bolalar nutqi statistik ma'lumotlar emas balki til o'zlashtirish jarayonini ochib beruvchi katta

ma'lumotlar (big data) sifatida qaralmoqda. L.Vygotsky aytganidek, "Zamonaviy tilshunoslikda bolalar nutqini o'rganish shunchaki lug'at boyligini sanash emas, balki miyadagi lisoniy mexanizmlarni shakllanish dinamikasini raqamli metodlar orqali tahlil qilishni anglatadi." [1]. Dunyo miqyosida bolalar nutqiga asoslangan korpus (CHILDES) ning yaratilishi, xalqaro tizimda ma'lumotlarni standartlashtirish nutq rivojlanishidagi tilga xos qonuniyatlarni aniqlashning ishonchli yo'liga aylandi. Ushbu korpus sabab zamonaviy ontolingvistikani raqamlashtirish imkoniyati paydo bo'ldi. "Bugungi kunda bolalar nutqini korpus asosida o'rganish nafaqat nazariy tilshunoslik, balki nutqiy patologiyalarni barvaqt aniqlash va sun'iy intellekt tizimlarini o'rgatishda ham "oltin standart" bo'lib xizmat qilmoqda" [2].

Hozirgi kunda o'zbek tilshunosligida bir qancha korpuslar mavjud. Masalan, N.Abduraxmonova tomonidan O'zbek tili milliy korpusi (<http://uzbekcorpus.uz/>) yaratilgan bo'lib, u o'z ichiga ko'plab badiiy, publitsistik va rasmiy hujjatlarini qamrab oladi. Biroq, mavjud korpuslar faqat kattalar nutqiga tayanadi. Demak, o'zbek tilshunosligida bolalar nutqi korpusi shakllanmagan va buni yaratish muhim vazifalardan biridir.

**Tadqiqot metodologiyasi.** O'zbek bolalar nutqi korpusini yaratish ko'p bosqichli metodologik yondashuvga asoslanadi. Xalqaro CHILDES (Child Language Data Exchange System) tizimiga muvofiq shuningdek, o'zbek tili agglyutinativ tabiati bilan hisoblashgan holda quyidagi bosqichlarda amalga oshiriladi:

Ma'lumotlar yig'ish va korpus dizayni. Tadqiqot materiali sifatida 3 yoshdan 7 yoshgacha bo'lgan o'zbekzabon bolalarning tabiiy muloqoti tanlab olindi. Ma'lumot yig'ishda tabiiy kuzatuv metodidan fodalaniladi. Masalan, bolaning kundalik o'yin jarayoni, oilaviy muloqoti, ertak yoki rasm asosidagi hikoya nutqi audio qurilmalar yordamida yozib olinadi.

Transkripsiya qilish jarayoni (Chat formati). Yozib olingan audio materiallar CHAT (Codes for the Human Analysis of Transcripts) asosida matnga ko'chiriladi. CHILDES asoschisi Brian MacWhinney aytganidek: "CHAT

Har bir nutqiy holat alohida qatordan yoziladi va muloqot ishtirokchilari quyidagicha belgilanadi:

Kod	To'liq nomi	Kim uchun ishlatiladi
*CHI:	Target child	Tadqiqot obyekti bo'lgan bola
*MOT:	Mother	Bolaning onasi
*FAT	Father	Bolaning otasi
*INV	Investigator	Tadqiqotchi
*GPA	Grandparent	Bobo yoki buvi
*SIB	Sibling	Bolaning akasi, opasi yoki uka, singlisi

Yuqorida aytib o'tganimizdek, o'zbek tili agglyutinativ tuzulishga ega bo'lgani uchun korpusdagi har bir so'z shakli

Morfema	vazifasi	chat(kodi)	Izoh
koptok	Asos	'n	koptok'
-cha	dimunitiv	-DIM	kichraytirish-erkalash
-lar	ko'plik	-PL	Plural(ko'plik)
-im	egalik	-PSSIS	Possessive 1st person singular (mening)
-ni	kelishik	-ACC	accusative (tushum kelishigi)

Ushbu jarayon korpusdagi eng muhim qismlardan biridir. U bolaning til tizimi qanchalik murakkab shakllanganini ko'rsatadi. CHILDESga ko'ra MLU (Mean Length of Utterance) ko'rsatkichi bolaning morfosintaktik kompetensiyasini ko'rsatuvchi jarayondir. "Gapning o'rtacha uzunligi (MLU) grammatik rivojlanishning ajoyib ko'rsatkichidir, chunki deyarli har qanday yangi bilim turi gapning uzunligini oshiradi"[4]. O'zbek tili korpusida esa MLUm (morfemalar bo'yicha uzunlik) qo'llanilishi maqsadga muvofiq "O'zbek tili agglyutinativ tabiatga ega bo'lganligi sababli, korpusda har bir so'z shakli morfemik darajada tahlil qilinishi shart"[5]. Masalan: Bola "kitoblarimni" so'zida bitta so'z aytgan bo'lsada uning miyasi 4ta ishni bajaradi: kitob(asos)+lar(ko'plik)+im(egalik)+ni(kelishik). Bu

Yoshi(age)	aniq yoziladi (yil;oy;kun)	Masalan:3:05:12 (3 yosh 5 oylik 12 kunlik)
Jinsi(sex)	female(qiz) male(o'g'il)	
Tili(language)	Bolaning ona tili va muloqot qiladigan boshqa tillar	masalan:uz, uz-ru

**Tahlil va natijalar.** O'zbek bolalar nutqi korpusini yaratish bo'yicha olib borilgan nazariy va metodologik izlanishlar quyidagicha natijalarni taqdim etadi:

Ma'lumotlarni yig'ish bosqichi. Bu bosqichda bolaning tabiiy muloqot muhiti audio yoki video formatda yozib olinadi. Bolaning faol lug'at boyligini aniqlash va grammatik shakllarni qo'llash mahoratini aniqlash uchun bolaning nutqi

tizimi yuzma-yuz muloqotdagi o'zaro ta'sirlarning kompyuterlashgan transkriptlarini yaratish uchun standartlashtirilgan formatni taqdim etadi"[3]. Bunda o'zbek tili uchun xos bo'lgan fonetik o'zgarishlar, pauzalar, takrorlar, nonverbal harakatlar maxsus belgilar bilan kodlanadi:

& – noto'g'ri fonetik shakl;

[/ ] – takrorlash;

[//] – o'z -o'zini tuzatish;

[/] – morfologik qayta qurish;

&=laughsk – kulgi;

&=cries – yig'i;

&=points:toy – o'yinchoqqa ishora qilish ;

&=gestures – turli imo-ishoralar;

xxx – tushunarsiz so'z;

yyy – tushunarsiz, lekin fonetik jihatdan so'zga o'xshash tovush;

+... – gapning chala qolishi.

morfemik tahlil qilinadi. Korpus quyidagi ikki bosqichli annotatsiyalanadi:

Leksik qatlam(lemma/stem)

Grammatik qatlam(Affixal tagging)

Birinchi bosqichda so'z barcha qo'shimchalardan tozalanadi va uning lug'aviy asosi aniqlanadi. Misol: Bolaning nutqidagi koptokchalarimni so'zi uchun leksik qatlam:n|koptok.

Ikkinchi bosqichda har bir qo'shimcha o'zining maxsus kodi bilan teglanadi. Bu o'zbek tili uchun muhim jarayon sanaladi chunki qo'shimchalar qat'iy ketma ketlikda keladi. Masalan:"koptokchalarimni"so'zi misolida buni qismlarga ajratamiz.

metod bola bir so'z tarkibida necha xil grammatik ma'noni birlashtira olishini aniqlab beradi.

Dasturiy ta'minot va tahliliy vositalar.Ma'lumotlarni qayta ishlash uchun ELAN va CLAN lingvistik dasturlaridan foydalaniladi:

1)CLAN(Computerized Language Analysis): Statistik tahlillarni amalga oshirish, so'zlarning chastotasini aniqlash va MLU, MLUm ko'rsatkichini avtomatik aniqlash uchun ishlatiladi.

2)ELAN(Eudico Linguistic Annotator); Video va audio fayllarni vaqt bo'yicha sigmentlash va ularni matnga ko'chirish(transkripsiya) uchun mo'ljallangan lingvistik dasturdir.Bolalar nutqini tahlil qilishda bu dastur muhim, chunki bolaning nafaqat ovozi balki harakatlari ham vaqt bilan bog'langan holda tahlil qilish imkonini beradi.

4.Etik tamoyillar:Tadqiqot jarayonida bolaning shaxsiy daxlsizligini ta'minlash uchun ota-onalardan ma'lumotlarni ilmiy maqsadda ishlatish uchun yozma rozilik olinadi. Korpusdagi bolalarning ismlari anonimlashtiriladi va raqamli identifikatorlar (masalan;Bola\_1) bilan almashtiriladi. Shuningdek, bola haqidagi quyidagi ma'lumotlar kiritiladi:

turli usullarda yozib olinadi. Masalan: Bolaga rasmga qarab hikoya qilish, ona va bola muloqoti kabi.

O'zbek tili uchun CHAT (transkripsiya modelini) ishlab chiqish. O'zbek tilining fonetik va morfologik xususiyatlarini hisobga olib chat tizimini standartlar asosida moslashtirishdir. Masalan: O'zbek tilidagi agglyutinativ

zanjirni tahlil qilish uchun morfemalar o'rtasida chegara belgisi bilan kodlanishi belgilandi(uy-i-miz-ni).

Morfologik teglash. O'zbek bolalar nutqi korpusini yaratishda morfemik annotatsiyalash tadqiqotning eng qiyin va

muhim qismlaridan biridir. Bunda har bir so'z morfemaga ajratiladi va teglanadi. Bunda so'z asosi va unga qo'shilgan affikslar bir-biridan ajratiladi. Masalan, kitobimizni so'zi quyidagicha kodlanadi:(n|kitob-PSS1P-ACC).

Turkum	Grammatik kategoriya	Teg	Misol
ot(noun)	ko'plik	-PL	koptok-lar(n koptok-PL)
	egalik(1-shaxs)	-PSS1S	uy-im(n uy-PSS1)
	egalik(2-shaxs)	-PSS2S	mushuk-ing(n mushuk-PSS2S)
	tushum kelishigi	-ACC	non-ni(n non-ACC)
	jo'nalish kelishigi	-DAT	maktab-ga(n maktab-DAT)
	o'rin payt kelishigi	-LOC	bog'-da(n bog'-LOC)
	chiqish kelishigi	-ABL	suv-dan(n suv-ABL)
Fe'l(verb)	o'tkan zamon	-PAST	kel-di(v kel-PAST.3S)
	hozirgi zamon	-PRES	bor-yapti(v bor-PRES.3S)
	buyruq mayli	-IMP	ber(v ber-IMP.2S)
	ravishdosh	-CONV	yugur-ib(v yugur-CONV)
	sifatdosh	-PART	uxla-yotgan(v uxla-PART)
Boshqa	kichraytirish	-DIM	gul-cha(n gul-DIM)

Avtomatik tahlil qilish. Korpus ma'lumotlari standartlarga keltirilgach, ularni tahlil qilish uchun CHILDES tizimining CLAN texnologiyasidan foydalaniladi. Ushbu tizim nutqiy ma'lumotlarni statistik va sifat jihatidan avtomatik tahlil qilishga yordam beradi. Bunda morfemalar soni MLUm orqali o'lchanadi (masalan:bor-ib kel-dim gapida 4ta morfema qatnashgan) hamda o'rtacha qiymat chiqaradi. Bu ko'rsatkich bolaning grammatik rivojlanish darajasini aniqlashga yordam beradi. R. Brauning ta'kidlashicha, "MLU shunchaki nutqning miqdoriy ko'rsatkichi emas, balki bolaning lisoniy tizimni qanchalik egallaganligini ko'rsatuvchi grammatik o'sish koeffitsiyentidir" [6].

Shuningdek, bu dasturda bolaning so'z boyligi TTR(Type-Token Ratio) aniqlanadi. Dastur bolaning nutqidagi umumiy(token) va noyob(type) so'zlarni aniqlaydi va ularni bir biriga nisbatlaydi.

**Xulosa va takliflar.** Tadqiqot natijalari o'zbek bolalar nutqini o'rganishda CHILDES tizimini qo'llashning bir necha jihatlarini ko'rsatadi. A. Aksu-Koçning fikricha, "Agglyutinativ tillarda har bir morfema qat'iy bir vazifada keladi va u boshqa qo'shimchalar bilan chalkashib ketmaydi"[7]. Buni o'zbek tili misolida tahlil qilsak. Masalan, ko'plik uchun -lar, jo'nalish kelishigi uchun -ga ishlatiladi bu esa bolaga tilning formulasini oson o'zlashtirib olishiga yordam beradi. Shuningdek, Dan Slobinning "Agglyutinativ tillarda barcha grammatik markerlar so'z oxiriga qo'shilgani sababli, bunday tilda so'zlashuvchi bolalar grammatikani boshqa tillarga qaraganda tezroq payqaydilar va o'zlashtiradilar"[8] kabi fikrlarini shuni ko'rsatdiki, o'zbek bolalari flektiv tillarga qaraganda murakkab morfemik birliklarni erta o'zlashtiradilar.

CHILDESga ko'ra bolalar nutqining o'rtacha uzunligi(MLU) muhim jarayon sanaladi. Agglyutinativ til sanaladigan o'zbek tili uchun esa qo'shimchalar soni(MLUm) ahamiyatlidir. R. Braun nazariyasiga ko'ra grammatik o'sish

jarayonini qo'shimchalar qo'llashda ko'rishimiz mumkin. Buni o'zbek bolalari nutqi misolida tahlilga tortsak: Masalan, bolaning "borolmayman" degan so'zi ingliz bolasining "I can not go" 4 ta so'ziga to'g'ri keladi. Bu jarayonda faqat so'zlarni sanasak ingliz bolasi grammatik rivojlanishda oldinda ko'rinishi mumkin. Lekin bu notog'ri xulosa. Mana shu joyda o'zbek tilining murakkabligi ko'zga tashlanadi. Borolmayman so'zini morfemalarga ajratamiz: Bor (asos/fe'l)+a (ravishdosh)+ol (ko'makchi fe'l)+ma (inkor)+y (ravishdosh)+man (shaxs-son). Bu yerda bola bitta so'z tarkibida 6ta morfemani birlashtirgan. Ingliz bolasining I+can+not+go gapida 4ta morfema bor. Bu jarayonni tahlil qilar ekanmiz o'zbek bolalar nutqini o'rganishda MLU o'lchovi yetarsizligini MLUm ni o'lchash ustuvorligini ko'rishimiz mumkin.

O'zbek bolalar nutqi korpusini yaratishning nazariy va metodologik asoslarini o'rganish quyidagicha xulosalarga kelishimizga asos bo'ldi:

– Xalqaro CHILDES tizimini o'zbek tili materiallarga moslashtirish milliy ontolingvistikamizni raqamli tahlil bosqichiga olib chiqadi. Shuningdek, tilimizni dunyo tillari bilan qiyoslab tahlil qilish uchun ham poydevor vazifasini o'taydi;

– Dunyodagi ontolingvistlar o'zbek tili misolida agglyutinativ tillar tabiatini yanada kengroq ko'ra oladilar. Shuningdek, o'z tadqiqotlarida tilimizning tabiatini tahlil qila boshlaydilar bu esa milliy tilimizning xalqaro mayqeyini oshiradi;

– Yaratilayotgan korpus nafaqat tilshunoslikda shuningdek, nutq nuqsonlarini o'rganadigan sohalar uchun ham manba vazifasini o'taydi;

– Shuningdek, sun'iy intellekt tizimlarini o'zbek bolalar nutqiga o'rgatishda muhim ahamiyat kasb etadi.

#### ADABIYOTLAR

1. Выготский Л.С. Мышление и речь. Собр. соч. в 6 т. Т. 2. М.: Педагогика, 1982. – С. 254-260.
2. MacWhinney, B. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. – Mahwah, NJ: Lawrence Erlbaum Associates, 2000. – P. 101-105.
3. MacWhinney, B. The CHILDES Project: Tools for Analyzing Talk. Volume 1: The CHAT Transcription Format. – 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates, 2000. – P. 15-18.
4. Brown, R.A First Language: The Early Stages. – Cambridge, MA: Harvard University Press, 1973. – P. 53-55.
5. Abdurahmonov G., Mahmudov N. O'zbek tili punktuatsiyasi va morfologiyasi masalalari. – Toshkent: Fan, 2011. – B. 88-92.
6. Miller, J.F. Assessing Language Production in Children: Experimental Procedures. – London: Edward Arnold, 1981. – P. 22-25.
7. Aksu-Koç, A.A., & Slobin, D. I. The acquisition of Turkish. In D. I. Slobin (Ed.), The crosslinguistic study of language acquisition. – Hillsdale, NJ: Erlbaum, 1985. – P. 839-844.

8. Slobin, D.I. Cognitive prerequisites for the development of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of child language development*. – New York: Holt, Rinehart & Winston, 1973. – P. 175-208.
9. Hamroyeva, Sh. O'zbek tili korpusi annotatsiya prinsiplari, 2021.P.– 42.
10. Abduraxmonova N. O'zbek tili elektron korpusining kompyuter modellari (monografiya), Globeedit, 2021. – 200 b.