INGLIZ TILI TRANSKRIPSIYALANGAN OG‘ZAKI KORPUSI UMUMIY TAVSIFI

Berdiyev Jahongir Botir o‘g‘li
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi mutaxassisligi 1-kurs magistranti

Mamasoyilova Sitora Mirzoxid qizi
1-kurs magistranti

Baxodirov Sanjarbek Rahmatali o‘g‘li
1-kurs magistranti

ANNOTATSIYA

Ushbu maqolada ingliz tili transkripsiyalangan og‘zaki korpusi, uning amaliy ahamiyati, qidiruv tizimi va imkoniyatlari, transkripsiyasi haqida umumiy ma’lumot berilgan. Amerika ingliz tilisining talabalar tomonidan transkripsiya qilingan korpusi og‘zaki tilning transkripsiyalari to‘plami bo‘lib, Amerika ingliz tilisidagi nutq shakllariga e’tibor qaratilgan. Ushbu korpus lingvistik tadqiqotlar, til o‘rganish va nutq texnologiyasini qo‘llashni rivojlantirishda qimmatli manba bo‘lib xizmat qiladi, og‘zaki tilni o‘rganish va tahlil qilishni muhim vositaga aylantiradi.

Kalit so‘zlar: og‘zaki korpus, vaziyat o‘zgaruvchilari, transkripsiya.

ABSTRACT

This article provides general information about English transcribed spoken corpus, its practical importance, search system and possibilities, and transcription. The Student Transcribed Corpus of American English is a collection of transcriptions of spoken language that focuses on American English speech patterns. This corpus is a valuable resource for the development of linguistic research, language learning, and speech technology applications, making it an important tool for the study and analysis of spoken language.

Key words: spoken corpus, situation variables, transcription.

KIRISH

Ingliz tilining talabalar tomonidan transkripsiyalangan og‘zaki korpusi talabalar tomonidan yaratilgan, yuqori sifatli nutq transkriptlari va ularga mos keladigan audio fayllar to‘plami hisoblanadi. Korpus intervyular, konferensiya suhbatlari va shaxsiy bloglar kabi turli xil sharoitlarda amerikalik ingliz tilida so‘zlashuvchilarning yozib olingan nutqlaridan namunalarni o‘z ichiga qamrab olgan. Korpus 2020-yilda ishga tushirilgan bo‘lsa ham, biroq uning uchun ma’lumotlar bazasi to‘plash ancha ilgari boshlangan. Korpusga bepul kirish va so‘rov natijalarini onlayn qidiruv interfeysi

orqali fayli ko‘rinishda yuklab olish mumkin. Korpusdan o‘qitish, tadqiqot yoki shunchaki qiziqish jihatidan foydalansa bo‘ladi.

ASOSIY QISM

Korpus hajmi 4-noyabr holatiga ko‘ra jami 176 511 ta so‘zshaklni qamrab oladi. 14 soat davom etgan nutqdan transkripsiya ishlari olib borilgan, 12 109 ta sintaktik jumla teglangan bo‘lsa, audio fayllarni yozib olishda 61 nafar talaba ishtirok etgan.

1. Qidiruv tizimi. Qidiruv tizimi korpusning asosiy bo‘limi hisoblanadi. Bu esa korpusning qay darajada tuzilganligini, ishlashini va imkoniyatlarini ko‘rsatib beradi. Korpusdan foydalanganda qidiruv natijalari KWIC(Key Word in Context) muvofiqlik chiziqlari sifatida ko‘rsatiladi.

Bu korpusning qidiruv interfeysi quyidagi xususiyatlarni o‘z ichiga oladi:

1. To‘liq so‘zlarni yoki so‘zlar ketma-ketligini qidirish mumkin:

“Dollar” bitta so‘z va bu so‘zning barcha misollarini topadi.” The atmosphere” birikma va buning ham barcha misollarini topadi.

2. Yulduzcha (*) belgisidan qidiruv kartasi sifatida foydlanish mumkin:

- wh* kabi qidirilganda wh bilan boshlangan barcha so‘zlarni topadi. Masalan, when, what;

- *ver* kabi qidirilganda tarkibida “ver” bo‘lgan so‘zlarni topadi. Masalan, version, never;

- th* w * kabi qidirilganda birinchi so‘zi th, ikkinchi so‘zi w bilan boshlangan so‘zlar ketma-ketligini topadi. Masalan, the water, things we;

- You * n’t kabi qidirilganda you va n’t orasida hech qanday so‘z bo‘lmagan misollarni topadi. Masalan, you can’t, you’re.

3. Nutqning bir qismi teglarini qidirish uchun pastki chiziqdan(_) foydalanish mumkin:

- _DT barcha aniqlovchilarni topadi. Masalan, the, a, all;

- _DT _NN aniqlovchilardan keyin kelgan birlikdagi otlarni topadi. Masalan, a person, this curve;

- Will_MD modal fe’l sifatida teglangan fe’llarni topadi. Masalan, what will they do;

- *ly_rb oxiri ly bilan tugaydigan fe’llarni topadi. Masalan, actually, really;

- _VV* VV bilan boshlanadigan nutq uchun teglangan barcha so‘zlarni topadi. Masalan, hozirgi zamon fe’li uchun VVP, look. O‘tgan zamon fe’li uchun esa VVP, called.

4. Lemmalarni qidirish uchun @ belgisidan foydalaniladi:

- @have lemmaning have shakllarini topadi. Masalan, have, had, has;

- @be_vvg lemmaning shakllarini topadi, keyin hozirgi zamon shakllari keladi. Masalan, is going, ‘m talking;

- _NN@RE* (yoki @RE*_NN). re bilan boshlangan birlik otlarni topadi. Masalan, reconstructions, reorganization.

5. So‘zlar, post-teglar yoki lemmalar ichida muqobillarini qidirish ushun | belgisidan foydalanish mumkin:

- heat|hot barcha issiqlik va issiq misollarini topadi;
- _VH*|VB* be yoki have fe’llari sifatida teglangan barcha shakllarni topadi.

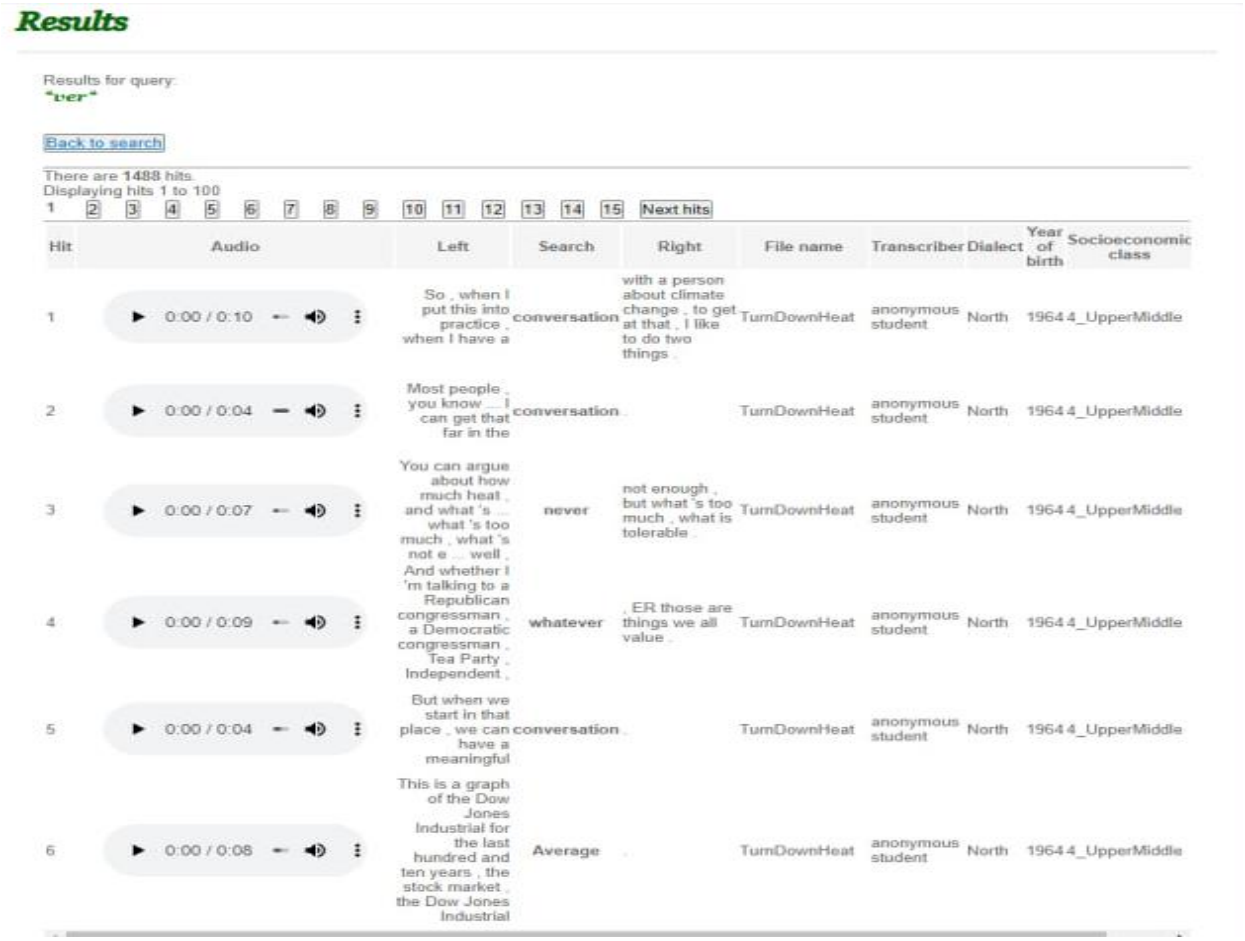
6. Qidiruv so‘zini ixtiyoriy qilish uchun dumaloq qavslar (…) dan foydalanish mumkin:

- _MD (not|n’t) _V* modal fe’ldan keyin kelgan barcha fe’llarni topadi va ikkalasining o‘rtasida inkor bo‘lishi mumkin. Masalan, can argue, may not have.

Masalan:

ver

qidiruv



1-rasm. Qidiruv natijasining ko‘rinishi

Qidiruv jarayonida:

Token identifikatori – token raqami, vaqti. Bu belgilar tanlanganda kiritilgan tokenning raqami va kiritilgan vaqti ko‘rsatiladi.

Fayl o‘zgaruvchilari – fayl raqami, fayl nomi, fayl manbayi, matn sarlavhasi, audio fayl qancha vaqt davom etishi, audio fayl necha soniyadan iborat ekanligi, faylda so‘zlar soni, tokenlar soni, yozib oluvchi, nutq egasining ismi, nutq manbasi kabi ma’lumotlarni olishimiz mumkin.

Hududi – joy nomi, davlat, kenglik, uzunlik, dialekt. Bu xususiyatlarni tanlaganimizda esa yozib olingan matnning qaysi hududga tegishli ekanligi yuqori darajada aniqlab bera oladi va o‘sha hududning dialektini ham ko‘rsatib beradi.

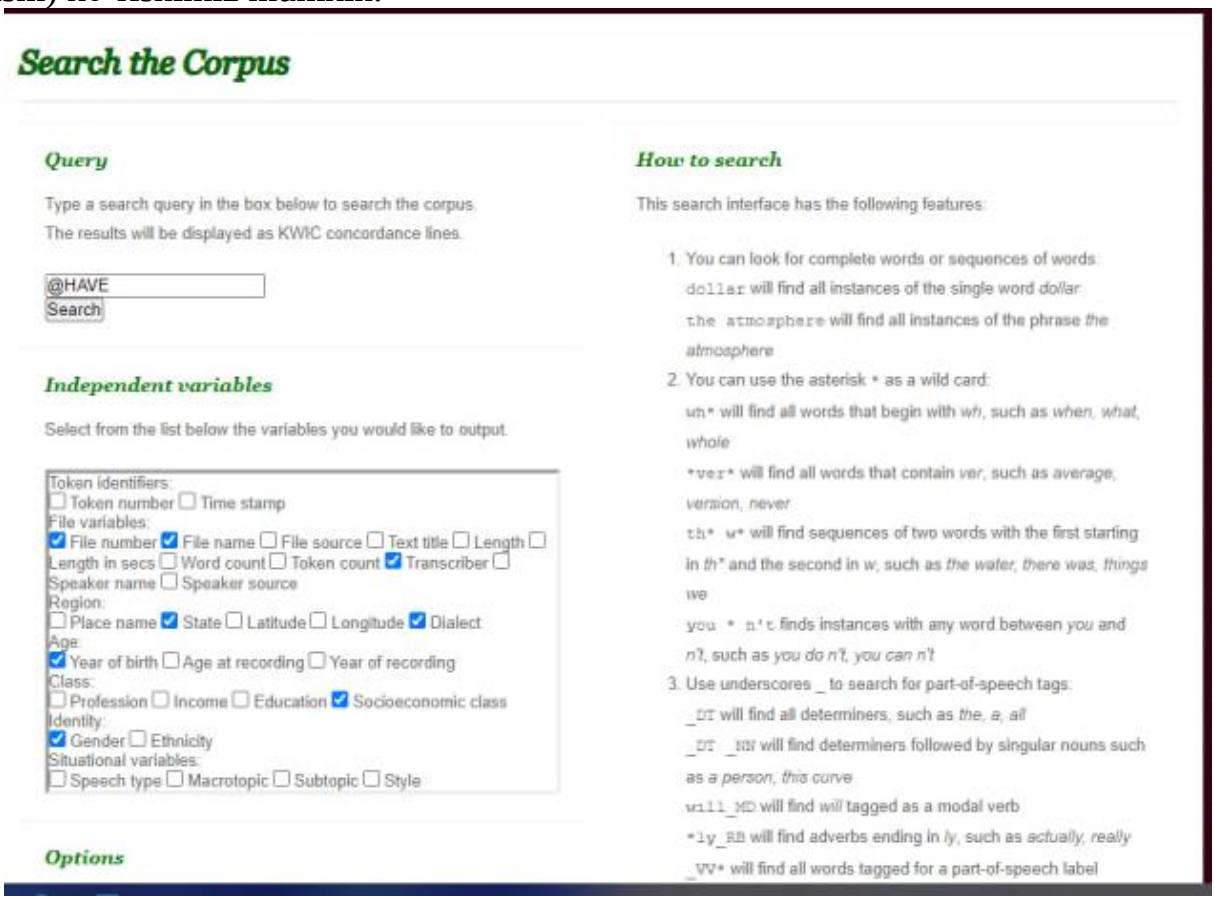
Yoshi – tug‘ilgan yili, yozib olish yoshi, yozilgan yili. Bu yerda nutq egasining tug‘ilgan yili, nutq yozib olinganda nutq egasi necha yoshda bo‘lgani, va yozilgan yili haqida ma’lumotlar olishimiz mumkin.

Sinfi – kasbi, daromadi, ta’lim darajasi, ijtimoiy sinfi kabi turi tanlanganda nutq egasining qaysi kasb egasi ekanligi, daromadi, ta’limning qaysi darajasiga ega ekanligi, va jamiyatda qaysi sinfga kirishi chiqarib beriladi.

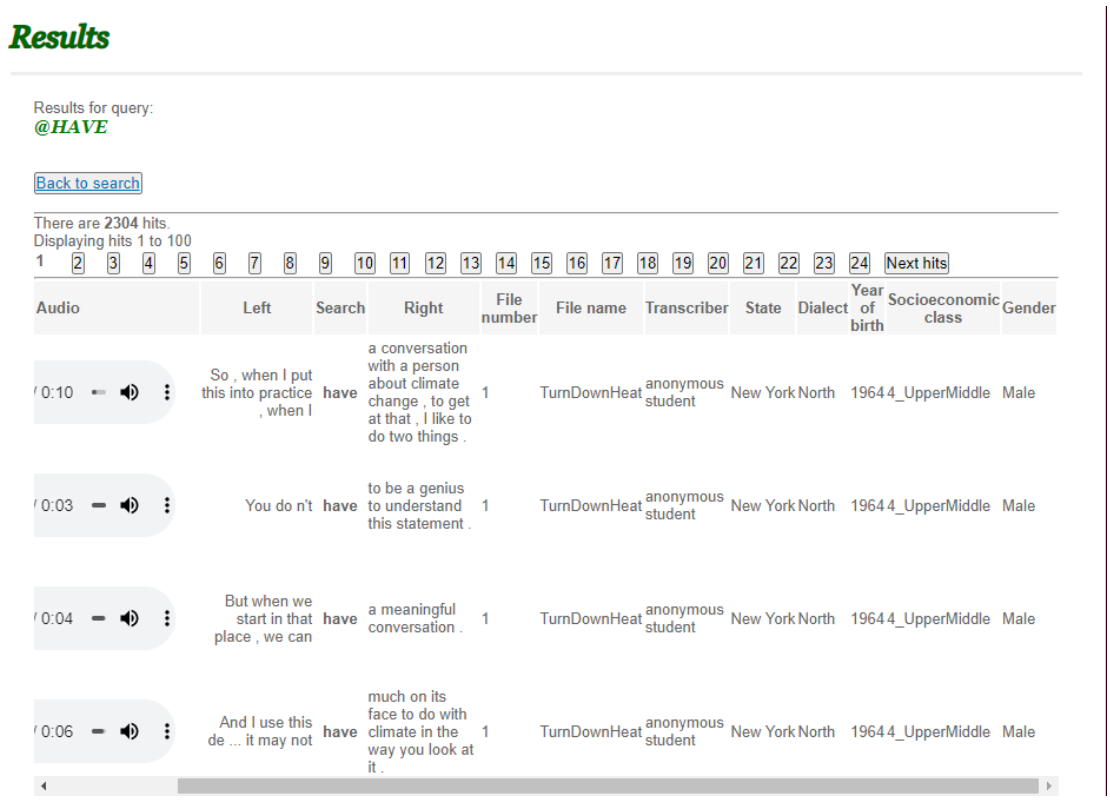
Jinsi bo‘limi tanlanganda nutq egasining qaysi jins vakili ekanligi aniqlanadi.

Vaziyat o‘zgaruvchilari – nutq turi, makrotopik, suhbat mavzusi, uslub(so‘zlovchining ruhiy holati) kabi belgilarni ko‘rish mumkin.

Quyida ba’zi o‘zgaruvchilarni tanlaganda (2-rasm) chiqarilgan natijani (3-rasm) ko‘rishimiz mumkin:



2-rasm. O‘zgaruvchilarni tanlash imkoniyati



Results

Results for query:
@HAVE

Back to search

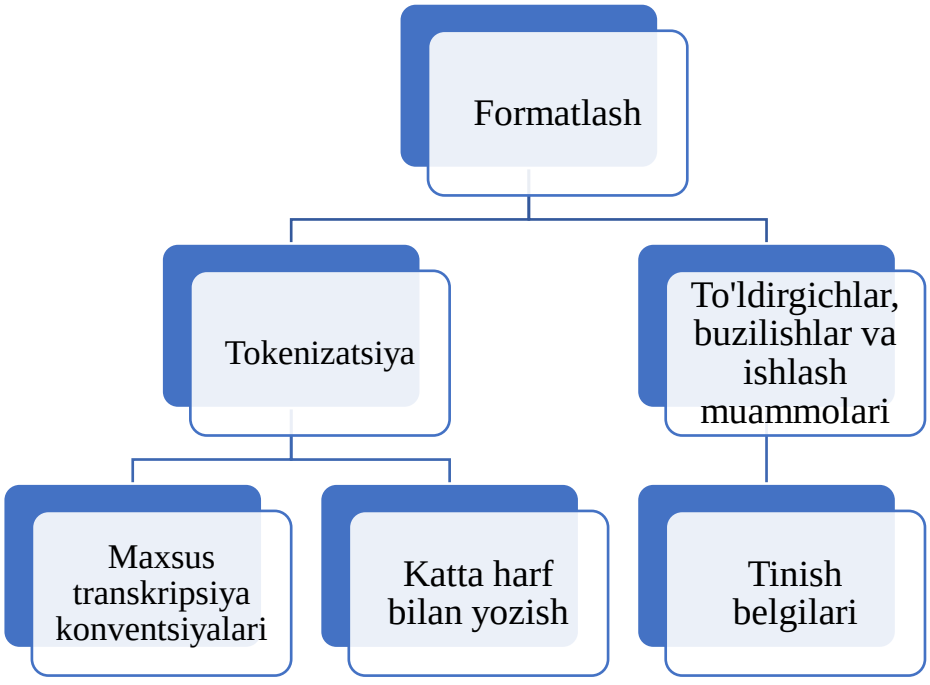
There are **2304** hits.
Displaying hits 1 to 100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 Next hits

Audio	Left	Search	Right	File number	File name	Transcriber	State	Dialect	Year of birth	Socioeconomic class	Gender
0:10	So , when I put this into practice , when I	have	a conversation with a person about climate change , to get at that , I like to do two things .	1	TurnDownHeat	anonymous student	New York	North	1964	4_UpperMiddle	Male
0:03	You do n't	have	to be a genius to understand this statement .	1	TurnDownHeat	anonymous student	New York	North	1964	4_UpperMiddle	Male
0:04	But when we start in that place , we can	have	a meaningful conversation .	1	TurnDownHeat	anonymous student	New York	North	1964	4_UpperMiddle	Male
0:06	And I use this de ... it may not	have	much on its face to do with climate in the way you look at it .	1	TurnDownHeat	anonymous student	New York	North	1964	4_UpperMiddle	Male

3-rasm. O‘zgaruvchilarni tanlaganda chiqarilgan natija

2. Transkripsiya

Ushbu sahifada korpus matn fayllarini yaratishda foydalaniladigan transkripsiya nazariyasining eng muhim jihatlari tushuntiriladi. Transkripsiya qoidalari izchillikni ta’minlash uchun iloji boricha ob’ektiv bo‘lishi uchun yaratilgan, lekin ayni paytda yaxshi o‘qilishi uchun nisbatan sodda. Tafsilotlar uchun transkripsiya bo‘yicha 1-jadvalga qarang:



1-jadval

2.1 . Vaqt belgilari va tokenizatsiya

Korpus to‘liq vaqt birliklari bilan qoplangan. Har bir yangi qurilma audiofaylning qayerda tegishli transkripsiya qilingan nutqni eshitish mumkinligini ko‘rsatadigan vaqt belgisiga ega. U [soat: daqiqa: soniya] formatiga ega, masalan, [00:03:12]. Korpus ham to‘liq jumla belgilariga tokenlashtirilgan. Gap belgisi mohiyatan mustaqil bosh gap bo‘lib, ochiq predmetli va chegaralangan fe’l hamda unga bog‘langan bo‘laklardan iborat. Tokenizatsiya transkripsiyaning eng qiyin jihatlaridan biridir. Noaniq holatlar, muvofiqlashtirish, qavs ichidagi bandlar va to‘g‘ridan-to‘g‘ri nutqdan tortib, umumiy ta’rifga qat’iy rioya qilmaydigan istisno belgilarigacha bo‘lgan juda ko‘p maxsus qoidalar mavjud. Vaqt birliklari bilan kiritilgan jumla belgilarining bir qator namunasi quyida ko‘rsatilgan, ega sariq rangda va fe’l yashil rangda ta’kidlangan.

[00:00:00] Welcome to this week’s "Top Stock Picks".
[00:00:02] I’m Tracey Ryniec.
[00:00:03] And I’m joined at the chairs this week by Sheraz Mian.
[00:00:06] And we have a couple of interesting stocks.
[00:00:09] One is an old Dow component.
[00:00:11] And the other one is a weight loss company, but not the one you might think.
[00:00:14] So Sheraz, we’re gonna start with you with the Dow component.
[00:00:18] I’m kind of surprised you picked Caterpillar because I haven’t been watching it but the last I looked, it was kind of down on its luck.

2.2. Imlo nutq funktsiyasining buzilishi

Nutqning chalkashligi – bu so‘zlovchi tomonidan yuzaga keladigan og‘zaki til oqimining har qanday buzilishi. Nutqdagi buzilishlarning turlariga duduqlanish va ikkilanishlar kiradi.Odatda yozma ravishda “ uhm , er , erm , uh “ h.k. sifatida berilgan umumiy to‘ldiruvchilar “ ER” sifatida bir xilda ko‘chiriladi (katta E, bosh R, tinish belgilari yo‘q). Quyidagi misollarda bu umumiy to‘ldiruvchi to‘q sariq rangda ko‘rsatilgan.

[00:00:09]	ER they’re doing a lot of the shell ER drilling up in the Dakotas, which is the really hot area right now.
[00:00:15]	And they’re seeing a lot of ER big finds up there.

Barcha turdagi buzilishlar uchta mustaqil nuqta bilan (...) ko‘rsatilgan. “Disfluens” – har xil turdagi parcha-parcha sintaktik birliklar. Ular uzoq yoki qisqa, murakkab yoki oddiy, takrorlashlar, tuzatishlar, baxtsiz hodisalar yoki noto‘g‘ri

boshlanishlar bo‘lishi mumkin. Quyidagi misollar uch nuqta bilan birga binafsha rangda keltirilgan materialni ko‘rsatadi.

[00:02:39]	But, you know, a ... their outlook is still very positive because they’re keeping their ... their costs in other areas.
[00:02:44]	And I don’t know if they have ever ... they’re doing quite well right now.

2.3. Bosh harf va tinish belgilari

Bosh harflar va tinish belgilari asosan ingliz tilidagi standart orfografiyaga mos keladi. Biroq, ba’zi farqlar ham mavjud. Masalan, to‘g‘ridan-to‘g‘ri nutq bitta vergulga kiritilgan. To‘g‘ridan-to‘g‘ri nutq quyida o‘qishda ko‘rsatilgan.

[00:20:10]	And I said, ‘Well, how much do you make?’.

Qo‘shiqlar, kitoblar va video o‘yinlar kabi media nomlari qo‘shtirnoq orasiga kiritilgan. Quyidagi jumlada kitob nomi ko‘k rangda ko‘rsatilgan:

[00:00:29]	If I were you, I’d write a book called "The Art of the Deal" because people are interested in deals.

Ko‘chirma matnda vergul, nuqta, so‘roq va tire, ba’zan esa erkin qo‘llaniladi. Boshqa tinish belgilari, masalan, nuqta-vergul, undov yoki qavslar umuman ko‘rinmaydi.

2.4. Izoh

Barcha korpus fayllari avtomatik ravishda teglangan va lemmatizatsiya qilingan. Foydalanuvchilar pastki chiziq bilan nutqning bir qismi teglarini (masalan, _DT barcha aniqlovchilarni topadi) va @ belgisi bilan lemmalarni qidirishlari mumkin (masalan, @TAKE “olish” lemmasining barcha so‘z shakllarini topadi).

2.5. Ogohlantirish

ANNOTATSIYA avtomatik ravishda amalga oshirilganligi va matnlar aytilganligi sababli - tragger o‘qitilmagan bir nechta orftografik konventsiyalarni o‘z ichiga olgan transkriptlar - avtomatik teglashning aniqligi unchalik yuqori bo‘lmasligi mumkin. Nutq qismi teglari yoki lemmalar yordamida qidiruvlar aniqlik va eslab qolish xatolariga olib kelishi mumkin. Avtomatik izohning ishlashi baholanmagan.

3. Yangiliklar

Og‘zaki Amerika ingliz tilining Talabalar tomonidan transkripsiyalangan korpusi ustida ish davom etmoqda. Bu yil yetti nafar bakalavriat talabalari 2023-yil bakalavriat stipendiyalari dasturi davomida korpusni yangi transkript bilan kengaytirdilar. Ular: Reet K Maur, Yan Li, Jorjina Uilobi, Yanxao Li, Kriti Mehrota, Faatima Adam va Greys Carrier. Ular 24 000 dan kam bo‘lmagan so‘zlarni qo‘shishga muvaffaq bo‘lishdi. Talabalarning yakuniy vitrinasi taqdimoti videosi hozirda mavjud. Ular

birinchi navbatda professional nutq korpusini yaratish haqida gapirishadi. Taqdimotni to‘g‘ridan-to‘g‘ri quyida ko‘rishingiz mumkin: https://youtu.be/KzCXqkJr4qY

XULOSA

Xulosa qilib aytganda, ingliz tilining transkripsiyalangan og‘zaki korpusi lingvistik tadqiqotlar, til o‘rganish va nutq texnologiyasini qo‘llashni rivojlantirish uchun muhim manba bo‘lib xizmat qiladi. U tildan foydalanish, dialektal variatsiyalar va nutq shakllari haqida qimmatli ma’lumotlarni taqdim etadi va ingliz tilining haqiqiy hayotiy kontekstlarda so‘zlashadigan keng qamrovli namunasini taqdim etadi. Bunday korpuslarning mavjudligi til haqidagi tushunchamizni rivojlantirish hamda og‘zaki nutqni to‘g‘rilash va ifodalashga qaratilgan vositalar va texnologiyalarni ishlab chiqish uchun zarurdir. Tilshunoslik va til texnologiyasi sohasi rivojlanishda davom etar ekan, yaxshi izohlangan, transkripsiyalangan og‘zaki nutqning ahamiyatini ularning til haqidagi bilimlarimizni oshirish va til bilan bog‘liq texnologiyalarni takomillashtirishdagi rolini oshirib bo‘lmaydi**.**

FOYDALANILGAN ADABIYOTLAR RO‘YXATI: (REFERENCES)

1. https://spokencorpus.org.
2. https://www.youtube.com/watch?v=8G6b9zdpuLU
3. https://www.youtube.com/watch?v=KzCXqkJr4qY