МРНТИ 16.31.21

З.Ә.Сиразитдинов[1], Л.А.Боҫконбаева[2],
Г.Г.Шәмсетдинова [3]

[1]Тарих, тел һәм әҙәбиәт институты әйҙәүсе ғилми хеҙмәткәре, ф.ф. к.
Өфө, Башҡортостан, РФ

[2]Тарих, тел һәм әҙәбиәт институты өлкән ғилми хеҙмәткәре, ф.ф.к.
Өфө, Башҡортостан, РФ

[3]Тарих, тел һәм әҙәбиәт институты кесе ғилми хеҙмәткәре.
Өфө, Башҡортостан, РФ

ДИАЛЕКТ КОРПУСЫ ӨСӨН АУДИОМАТЕРИАЛДАРҘЫ ЙЫЙЫУ ҺӘМ ЭШКӘРТЕҮ МӘСЬӘЛӘЛӘРЕ

Аннотация. Мәҡәләлә башҡорт теленең көнсығыш диалектының аудиоматериалдар корпусын төҙөү проблемаһы күтәрелә. Мәҡәләлә корпуста бирелгән метамәғлүмәт (информанттың йәше, белем кимәле, ғаиләлә аралашыу теле, милләте һ.б.) ентекләп ҡарала. Темалар төрлөлөгө диалекттың фонетик, лексик һәм грамматик үҙенсәлектәрен сағылдырырға мөмкинлек бирә. Корпуста аудиофайлдарҙың транскрипцияһы, әҙәби варианты, урысҫаға тәржемәһе урын аласаҡ.

Терәк һүҙҙәр: башҡорт теле, төрки телдәр, корпус лингвистикаһы, диалект корпус, экспедиция материалдары.

З.Ә.Сиразитдинов[1], Л.А.Босқонбаева[2],
Г.Г.Шәмсетдинова [3]

[1]РҒА тарих, тіл және әдебиет институтының жетекші ғылыми қызметкері, ф.ғ.к.
Уфа, Башқұртстан, РФ
[2]РҒА тарих, тіл және әдебиет институтының аға ғылыми қызметкері, ф.ғ.к.
Уфа, Башқұртстан, РФ

[3]РҒА тарих, тіл және әдебиет институтының кіші ғылыми қызметкері.
Уфа, Башқұртстан, РФ

ДИАЛЕКТІЛІК КОРПУС ҮШІН АУДИОМАТЕРИАЛДАРДЫ ЖИНАУ ЖӘНЕ ӨҢДЕУ МӘСЕЛЕЛЕРІ

Аннотация. Мақалада башқұрт тілінің Шығыс диалектісіне арналған аудиоматериалдар корпусын құру мәселесі көтеріледі. Мақалада халық тілінен жинақталған деректер (жасы, мектептегі ақпарат деңгейі, қарым-қатынас тілі, ұлты және т.б.) мұқият қарастырылады. Диалектінің гетерогенділігі диалектінің фонетикалық, лексикалық және грамматикалық ерекшеліктерін көрсетуге мүмкіндік береді. Корпуста аудио файлдарға арналған транскрипция, әдеби нұсқа, орыс тіліндегі аударма болады.

Тірек сөздер: башқұрт тілі, түркі тілдері, корпус лингвистикасы, диалект корпусы, экспедиция материалдары.

**З.А.Сиразитдинов[1], Л.А.Босконбаева[2],
Г.Г.Шамсетдинова [3]**

[1]Ведущий научный сотрудник Института истории, языка и литературы РАН, к.ф.н.
Уфа, Башкортостан, РФ

[2]Старший научный сотрудник Института истории, языка и литературы РАН, к.ф.н.
Уфа, Башкортостан, РФ

[3]Младший научный сотрудник Института истории, языка и литературы РАН.
Уфа, Башкортостан, РФ

ВОПРОСЫ СБОРА И ОБРАБОТКИ АУДИОМАТЕРИАЛОВ ДЛЯ ДИАЛЕКТНОГО КОРПУСА

Аннотация. В настоящей статье рассматриваются прицитпы создания корпуса аудиоматериалов восточного диалекта башкирского языка. Перед авторами ставится задача репрезентативного сбора полевых материалов с учетом возрастной, гендерной принадлежности, уровня образования, языка общения, национальной идентичности информантов и разнообразия тем для беседы. В корпусе предполагается представление аудиозаписи, расшифровка звуковых файлов в виде транскрибированных текстов, их литературный вариант и русский перевод диалектного текста.

Ключевые слова: башкирский язык, тюркские языки, корпусная лингвистика, диалектный корпус, экспедиционный материал.

**Z. A. Sirazitdinov [1], L. A. Boskonbayeva [2],
G. G. Shamsetdinova [3]**

[1] Leading Researcher of the Institute of History, Language and Literature of the Russian Academy of Sciences, candidate of philological sciences,
Ufa, Bashkortostan, Russia

[2] Senior Researcher of the Institute of History, Language and Literature of the Russian Academy of Sciences, candidate of philological sciences,
Ufa, Bashkortostan, Russia

[3] Junior Researcher of the Institute of History, Language and Literature of the Russian Academy of Sciences.
Ufa, Bashkortostan, Russia

ISSUES OF COLLECTING AND PROCESSING AUDIO MATERIALS FOR THE DIALECT CORPUS

Annotation. This article discusses the principles of creating a corpus of audio materials of the eastern dialect of the Bashkir language. The authors are tasked with representative collecting field materials, taking into account the age, gender, level of education, language of communication, national identity of informants and the variety of topics for conversation. The corpus is intended to present audio recordings, to transcribe audio files as transcribed texts, to present their literary version and the Russian translation of the dialect text.

Keywords: the Bashkir language, Turkic languages, corpus linguistics, dialect corpus, expedition material.

Йәнле һөйләү телен өйрәнеүҙә, халыҡтың һүҙ байлығын туплауҙа корпустарҙың әһәмиәтен һәм ролен күҙ уңында тотоп, башҡорт теленең диалекттары һәм һөйләштәре буйынса репрезентатив корпус төҙөү телселәр алдына ҡуйылған төп бурыстарыҙың береһе булып тора. Был корпуста башҡорт теленең барлыҡ һөйләштәренең өлгөләре лә урын алырға тейеш. Информанттың йәшен, белем кимәлен, ғаиләлә аралашыу телен, милләтен күрһәткән метамәғлүмәттең булыуы, темаларҙың төрлөлөгө диалекттың фонетик, лексик һәм грамматик үҙенсәлектәрен сағылдырырға мөмкинлек биргән был корпус төрлө социологик тикшеренеүҙәр үткәреү өсөн дә ҙур әһәмиәткә эйә буласаҡ. Темаларҙың төрлөлөгө һөйләштең фонетик, лексик һәм грамматик үҙенсәлектәрен сағылдырырға мөмкинлек бирәсәк. Корпуста аудиофайлдарҙың расшифровкаһы, әҙәби варианты, урыҫсаға тәржемәһе урын аласаҡ.

Диалект һәм һөйләш үҙенсәлектәрен экспедиция материалдары нигеҙендә өйрәнеү диалектологик тикшеренеүҙәр өсөн яңы мөмкинлектәр аса, һөйләштәрҙең сиктәрен билдәләргә, уларҙың фонетик, лексик һәм грамматик үҙенсәлектәрен асыр-ға булышлыҡ итә.

Башҡорт телен һәм уның диалекттарын комплекслы өйрәнеү XX быуаттың 20–30-сы йылдарында башлана. Диалекттарҙы тикшереүгә башҡорт ғалимдары Ғ.С. Амантаев, Т. Ғ.Байышев, Ғ.Й. Дәүләтшин һәм З.Ш. Шәкировтың индергән өлөшө ҙур. Улар күп һанлы экспедициялар һөҙәмтәһендә диалекттарҙың һәм һөйләштәрҙең сиктәрен, уларҙың төп үҙенсәлектәрен билдәләүгә ирешә. Өҫтә һанап үтелгән ғалимдарҙы, һис шикһеҙ, башҡорт диалектологияһы фәненә нигеҙ һалыусылар тип әйтергә була.

50-се йылдарҙа телсе ғалимдар ерле һөйләштәрҙең фонетикаһын, лексикаһын һәм морфологияһын системалы тикшереүгә күсә. Т.Ғ. Байышевтың, Ж.Ғ. Кейекбаевтың был өлкәләрҙә өйрәнеүгә арналған монографик хеҙмәттәре донъя күрә [1-2]. Уларҙа ғалимдар диалекттарҙы, һөйләштәрҙе тасуирлап, таралыу сиктәрен билдәләп кенә ҡалмай, телдең фонетик, лексик һәм морфологик кимәлдәрендә әҙәби тел һәм диалекттарҙың үҙ-ара мөнәсәбәт проблемаларын да тикшерә. 60–70-се йылдарҙа үткәрелгән экспедициялар ваҡытында тупланған махсус материал диалект лексикографияһына нигеҙ һала [3-5], диссертацион һәм монографик хеҙмәттәр [6-9] өсөн ҡиммәтле сығанаҡ булып тора.

1973 йылда башҡорт диалектологтары Башҡорт теленең диалектологик атласы өҫтөндә эш башлай. Башҡорт һөйләштәре лингвогеографик ысул менән тикшерелеп, һәр бер территориаль тел үҙенсәлеге картаға төшөрөлә. Изоглоссалар теге йәки был күренештең географик таралышын ғына күрһәтеп ҡалмай, шулай уҡ уның характерын, үҫеш юлын күрһәткән белешмә лә бирә. Ҡыҙғаныҫҡа ҡаршы, был фундаменталь хеҙмәт тик 2005 йылда ғына баҫылып сыға [10]. Атлас материалдары Башҡорт теленең машина фондының диалектологик базаһына индерелгән [11].

80-се йылдарҙан алып бөгөнгө көндәргәсә башҡорт диалекттары һәм һөйләштәре фронталь тикшеренеүҙәрҙән ситтә ҡала килә. Ялан шарттарында махсус материал йыйыу маҡсатында берҙәм программа буйынса ойошторолған экспедицияларҙың һаны ҡырҡа кәмей. Экспедиция материалдарын йыйғанда информанттарҙың йәше, милләте, белеме һәм башҡа социаль үҙенсәлектәре иҫәпкә алынмай. Элекке экспедицияларҙа йыйылған материалдарҙың күп өлөшө заманса форматҡа күсерелмәгән йәиһә бөтөнләй юғалып ҡалған.

Һис шикһеҙ, 30 йыл эсендә диалекттарҙа һәм һөйләштәрҙә төрлө үҙгәрештәр булған. Әҙәби тел йоғонтоһо аҫтында күп диалектизмдар юғала бара. Шуға күрә лә

уларҙы бер урынға туплау ғына түгел, ғилми яҡтан анализлау ҙа талап ителәә. Бөгөнгө лингвистика һөйләштәрҙең телен тикшереүгә яңы ысулдар һәм алымдар тәҡдим итә. Был ысулдар тел материалдарын тиҙ һәм объектив рәүештә эшкәртергә ярҙам итә.

Һуңғы йылдарҙа корпус лингвистикаһы тел ғилеме өлкәһендә әүҙем үҫешкән йүнәлештәрҙең береһе булып тора. Ул XX быуаттың икенсе яртыһында ҙур күләмле текстарҙы компьютерҙа эшкәртеү һөҙөмтәһендә барлыҡҡа килеп, хәҙер инде тел ғилеменең айырым тармағына әүерелде.

Корпус лингвистикаһының объекты – һөйләү һәм яҙма телмәр материалдары (текстар). Донъяның бөтә эре телдәре үҙҙәренең милли корпустарын булдырған. Уларҙың һаны йылдан-йыл арта ғына бара һәм әлеге көндә меңгә яҡын лингвистик корпус булыуы билдәле.

Корпус лингвистикаһының ысулдары һәм алымдары диалект материалдарын үҙ эсенә алған корпустарҙы төҙөүҙә лә киң ҡулланыла. Диалект корпустарының бер өлөшө милли корпустар составына индерелһә, икенсе өлөшө үҙаллы корпус булараҡ төҙөлгән. Мәҫәлән, Урыҫ теленең милли корпусы 12 корпустан тора, улар араһында диалект корпусы ла урын алған.

Әлеге көндә Рәсәйҙә диалект корпустарын төҙөү эше әүҙем алып барыла. Коми теленең ижем һөйләше корпусы, Урыҫ һөйләштәренең Һарытау корпусы, Архангель өлкәһе Устьян районының диалект текстарының мультимедия корпусы, Урыҫ теленең милли корпусына ҡараған диалект корпусы булдырылған һәм Интернет селтәренә ҡуйылған.

Башҡорт диалекттарының корпусын төҙөү мәсьәләһе башҡорт тел ғилемендә беренсе тапҡыр күтәрелә. Шул уҡ ваҡытта коллективтың башҡорт теленең лингвистик корпустарын төҙөүҙә бай тәжрибәһе бар. Әлеге көндә Интернет селтәрендә башҡорт теленең проза, ваҡытлы матбуғат һәм халыҡ ижады текстары буйынса төҙөлгән электрон корпустары эшләп килә [12-14].

Йәнле һөйләү телен өйрәнеүҙә, халыҡтың һүҙ байлығын туплауҙа корпустарҙың әһәмиәтен һәм ролен күҙ уңында тотоп, башҡорт теленең диалекттары һәм һөйләштәре буйынса репрезентатив корпус төҙөү телселәр алдына ҡуйылған төп бурыстарыҙың береһе булып тора. Был корпуста башҡорт теленең барлыҡ һөйләштәренең өлгөләре лә урын алырға тейеш. Унда информанттың йәшен, белем кимәлен, ғаиләлә аралашыу телен, милләтен күрһәткән метамәғлүмәттең булыуы төрлө социологик тикшеренеүҙәр үткәреү өсөн дә ҙур әһәмиәткә эйә буласаҡ. Темаларҙың төрлөлөгө һөйләштең фонетик, лексик һәм грамматик үҙенсәлектәрен сағылдырырға мөмкинлек бирәсәк. Корпуста аудиофайлдарҙың расшифровкаһы, әҙәби варианты, урыҫсаға тәржемәһе урын аласаҡ.

Лингвистика һәм мәғлүмәт технологиялары лабораторияһы коллективы проекттың әлеге этабында башҡорт теленең көнсығыш диалекты буйынса телмәр материалын йыйыу һәм уны эшкәртеү буйынса эш алып бара. Был диалект әй, һалйот, арғаяш, миәс, ҡыҙыл, учалы һөйләштәрен үҙ эсенә ала.

1) Көнсығыш диалекттың фонетикаһы, лексикаһы, морфологияһы һәм синтаксисының территориаль үҙенсәлектәрен сағылдырған махсус материал йыйыу өсөн Башҡортостан Республикаһына, Силәбе, Свердловск һәм Ҡурған өлкәләренә ҡараған 28 торама пункт һайлап алынды. Улар шулай уҡ Башҡорт теленең диалектологик атласы өсөн дә терәк пункт итеп алынған.

2) Һәр бер торама пунктта кәмендә 12 информант яҙып алына: гендер буйынса 2 яҙма, йәш төркөмө буйынса 6 яҙма. Түбәндәге йәш төркөмдәре айырыла:

– мәктәпкәсә йәштәге һәм башланғыс класс уҡыусылары (11 йәшкә тиклем),

– урта йәштәге мәктәп уҡыусылары (11–15 йәш),

– өлкән йәштәге мәктәп уҡыусылары һәм студенттар (16–25 йәш),

– урта йәштәгеләр (25–45 йәш),

– өлкән йәштәгеләр (45–65 йәш),

– оло йәштәгеләр (65 йәштән алып).

Мәғлүмәт биреүсенең белем кимәле (башланғыс, урта, юғары) иҫәпкә алына.

3) Аудиояҙма менән бергә файлдың экстралингвистик билдәләрен үҙ эсенә алған мәғлүмәт яҙып алына:

– зат: ир-ат, ҡатын-ҡыҙ;

– белем: башланғыс, урта (урта мәктәп йәки урта профессиональ белем биреү учреждениеһы), юғары;

– йәш;

– белем алыу теле: башҡорт, урыҫ, татар, сыуаш һ.б.;

– ғаиләлә аралашҡан теле: башҡорт, урыҫ, татар;

– мәғлүмәт биреүсенең милләте;

– исеме, атаһының исеме, фамилияһы;

– йәшәгән урыны;

– был урынға күскәнсе һуңғы йәшәү урыны (күскән осраҡта); – яҙыу ваҡыты.

4) Монолог, диалог һәм полилог кеүек аралашыу төрҙәре айырыла.

5) Информанттарҙан мәғлүмәт яҙып алыу өсөн 15 тема билдәләнгән:

– туй, туй йолалары;

– ашамлыҡ (көндәлек аштар, ҡунаҡ аштары);

– мал-тыуар (ниндәй мал-тыуар аҫрайҙар, нисек ҡарайҙар);

– туғанлыҡ системаһы (балалар һәм яҡын туғандар);

– өй янындағы биләмәләр (емеш, йәшелсә баҡсаһы);

– йәшәгән йорт (ҡасан һалынған, кем тарафынан һалынған);

– ауыл тирәһендәге ер-һыу атамалары;

– ауыл, мәктәп, ырыу тарихы;

– көндәлек тормош (эш, мәктәп, дуҫтар тураһында әңгәмә);

– миҙгел, һауа торошо;

– таҡмаҡтар, мәҡәл-әйтемдәр, әкиәттәр;

– район үҙәгенә, күрше ауылдарға барып ҡайтыу (ниндәй эштәр буйынса, ниндәй транспорт менән);

– балаларҙың уйындары (ниндәй уйындар һәм нисек уйнайҙар);

– балаларҙың дуҫтары (кемдәр улар, ҡайҙа йәшәйҙәр);

– ауыл тирәһендә осраған хайуандар (ниндәй хайуандар һәм ҡоштар осрай).

6) Яҙма цифрлы диктофонға ҡыҫылмаған форматта (.WAV, 16бит/22kHz – 16бит/48kHz) яҙыла. Әңгәмә барышына ҡамасаулаған шау-шыу, тауышты (телевизор, радио һ.б.) туҡтатып тороу һорала. Аудиояҙма Sound Forge программаһы ярҙамында паспортизация үтә һәм төрлө шау-шыуҙан һәм оҙон паузаларҙан таҙартыла.

7) Файлдарҙың паспортизацияһы һәм метамәғлүмәттәр аудиофайлдың исеменә индерелгән:

1. v – к өнсығыш диалект, aj – әй һөйләше, ar – арғаяш һөйләше, sl – һалйот һөйләше, mi – миәс һөйләше, kz – ҡыҙыл һөйләше; uc – учалы һөйләше;

2. m – ир-ат, w – ҡатын-ҡыҙ;

3. белем: n – башланғыс белем, s – урта белем, v – юғары белем;

4. йәшкә бәйле төркөмдәр номерҙары: 1 – мәктәпкәсә йәштәге һәм башланғыс класс уҡыусылары (11 йәшкә тиклем), 2 – урта йәштәге мәктәп уҡыусылары (11–15 йәш), 3 – өлкән йәштәге мәктәп уҡыусылары һәм студенттар (16–25 йәш), 4 – урта йәштәгеләр (25– 45 йәш), 5 – оло йәштәгеләр (45–65 йәш), 6 – өлкән йәштәгеләр (65 йәштән алып);

5. t1 – t15 – темалар;

Мәҫәлән, varwvt14d003 файлы исеменән көнсығыш диалектының арғаяш һөйләшендә һөйләшкән урта йәштәге юғары белемле ҡатын-ҡыҙҙан “туй, туй йолалары” темаһына яҙып алынған аудиофайл икәнлеген белергә була. Информанттың өҫтәлмә кодтары d001 алып d999 тиклем. Был код текст файлының исеме булып тора. Унда мәғлүмәт биреүсенең исеме, атаһының исеме, фамилияһы, уның ғаиләлә аралашҡан теле, милләте, әлеге йәшәгән урынға тиклем йәшәү урыны (күскән осраҡта), был урында йәшәү ваҡыты, яҙыу датаһы бирелә. Һәр бер мәғлүмәт @ билдәһе менән айырыла. Мәҫәлән, Зилдә Әхмәҙи ҡыҙы Мәүлитова@б@б@б@Ташҡыя Учалы районы БР@10@25@12.08.2017, был файлдың исеме – d005.txt.

Әлеге көндә 45 мәғлүмәт биреүсенең материалдары эшкәртелеп транскрипцияланды. Улар башҡорт теленең диалект корпусы һөйләштәрҙең фонетикаһын, лексикаһын һәм грамматикаһын өйрәнеүҙә баһалап бөткөһөҙ ҡиммәтле сығанаҡ буласаҡ.

Әҙәбиәт исемлеге:

[1] Баишев Т.Г. Башкирские диалекты в их отношении к литературному языку. – Уфа, 1953. – 132 с.

[2] Киекбаев Дж.Г. Башкирские диалекты и краткое введение в их историю. – Уфа, 1958.

[3] Башҡорт һөйләштәренең һүҙлеге. 1 т. Көнсығыш диалект / Н.Х.Мәксүтова һ.б. – Өфө, 1967. – 300 б.

[4] Башҡорт һөйләштәренең һүҙлеге. 2 т. Көньяҡ диалект / Н.Х.Мәксүтова һ.б. – Өфө, 1970. – 327 б.

[5] Башҡорт һөйләштәренең һүҙлеге. 3 т. Көнбайыш диалект / Н.Х.Мәксүтова һ.б. – Өфө, 1987. – 231 б.

[6] Ишбулатов Н.Х. Диалектная система башкирского языка в сравнительно-историческом освещении: Автореф. дисс. д.ф.н. – Уфа, 1974.

[7] Максютова Н. Х. Восточный диалект башкирского языка в сравнительно-историческом освещении. – М.: Наука, 1976. 292с.

[8] Миржанова С. Ф. Южный диалект башкирского языка. – М.: Наука, 1979. 272с.

[9] Миржанова С.Ф. Северо-западный диалект башкирского языка. – Уфа, 1991. 296 с.

[10] Диалектологический атлас башкирского языка. – Уфа, 2005. – 232 с.

[11] Машинный фонд башкирского языка. URL:http://mfbl2.ru (дата обращения: 25.10.2020).

[12] Сиразитдинов З.А., Бускунбаева Л.А., Бардыбаева А.Д., Ишмухаметова А.Ш., Рафикова Л.Н. О разработке национального корпуса башкирского языка // Роль и место национальной библиотеки в социокультурном пространстве. Материалы Международной научно-практической конференции. – Уфа, 2012. – С. 118-126.

[13] Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г. Корпус текстов периодической печати на башкирском языке // Актуальные проблемы диалектологии языков народов России. – Уфа, 2012. – С. 139-141.

[14] Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. О создании корпуса башкирского фольклора // Урал-Алтай: через века в будущее. Материалы VI Всероссийской тюркологической конференции (с международным участием). – Уфа, 2014. – С. 86-89.