

Торайғыров университетінің
ҒЫЛЫМИ ЖУРНАЛЫ

НАУЧНЫЙ ЖУРНАЛ
Торайғыров университета

**ТОРАЙҒЫРОВ
УНИВЕРСИТЕТІНІҢ
ХАБАРШЫСЫ**

Филологиялық серия
1997 жылдан бастап шығады



**ВЕСТНИК
ТОРАЙҒЫРОВ
УНИВЕРСИТЕТА**

Филологическая серия
Издается с 1997 года

ISSN 2710-3528

№ 1 (2023)

Павлодар

**НАУЧНЫЙ ЖУРНАЛ
ТОРАЙГЫРОВ УНИВЕРСИТЕТА**

Филологическая серия
выходит 4 раза в год

СВИДЕТЕЛЬСТВО

О постановке на переучет периодического печатного издания,
информационного агентства и сетевого издания

№ KZ30VPY00029268

выдано

Министерством информации и общественного развития
Республики Казахстан

Тематическая направленность
публикация материалов в области филологии

Подписной индекс – 76132

<https://doi.org/10.48081/ICYL6092>

Бас редакторы – главный редактор

Жусупов Н. К.

д.ф.н., профессор

Заместитель главного редактора

Анесова А. Ж., *доктор PhD*

Ответственный секретарь

Уайханова М. А., *доктор PhD*

Редакция алқасы – Редакционная коллегия

Дементьев В. В., *д.ф.н., профессор (Российская Федерация)*

Еспенбетов А. С., *д.ф.н., профессор*

Трушев А. К., *д.ф.н., профессор*

Маслова В. А., *д.ф.н., профессор (Белоруссия)*

Пименова М. В., *д.ф.н., профессор (Российская Федерация)*

Баратова М. Н., *д.ф.н., профессор*

Аймухамбет Ж. А., *д.ф.н., профессор*

Шапауов Ә. Қ., *к.ф.н., профессор*

Шокубаева З. Ж., *технический редактор*

За достоверность материалов и рекламы ответственность несут авторы и рекламодатели

Редакция оставляет за собой право на отклонение материалов

При использовании материалов журнала ссылка на «Вестник Торайгыров университета» обязательна

<https://doi.org/10.48081/PABZ8209>

***Г. О. Сакиева¹, Л. Ә. Ибраймова², Б. Молдағали³**

^{1,2}Қазақ ұлттық қыздар педагогикалық университеті,

Қазақстан Республикасы, Алматы қ.;

³Абай атындағы Қазақ ұлттық педагогикалық университеті,

Қазақстан Республикасы, Алматы қ.

*e-mail: gulzirasakieva49@gmail.com

МУЛЬТИМЕДИЯЛЫҚ КОРПУС ЖАСАУДЫҢ ПРОБЛЕМАЛАРЫ

Қазақ тілінің мультимедиялық корпусын жасау – көп зерттеуді қажет ететін тың сала. Ауызекі сойлеу тілі бүгінгі күнге дейін тек мәтінді салыстыру, талдау деңгейінде ғана қарастырылып келді. Мультимедиялық корпус екінші тіл ретінде қазақ тілін меңгертуде өте тиімді құрал бола алады. Ауызекі сойлеу тілі мен әдеби тіл нормасындағы ауытқулар мен сәйкессіздіктерді анықтауда зор үлесін қоса алады. Бұл мақалада мультимедиялық корпус жасау барысында туындайтын проблемалар анықталды. Проблемаларды анықтау – корпус жасау жұмысын жеңілдетіп, алдын ала кедергілерді жеңуге көмегін тигізеді. Ол үшін «Корпус» терминіне қысқаша түсініктеме беріліп, соның ішіндегі «мультимедиялық корпус» анықталды. Қазақ тілінің ұлттық корпусы және шетелдік НКРЯ (Орыс тілінің ұлттық корпусы) құрамына енетін мультимедиялық корпусының (МУРКО) мысалында қысқаша сипаттама жасалды. Осы саланы зерттеген ғалымдардың тұжырымдарына сүйеніп, мультимедиялық корпусстардың маңыздылығы және қажеттілігі айқындалды. Сонымен қатар корпуссты жасау барысында туындайтын қиындықтар қарастырылды. Олар бірнеше түрге бөлінді. Атап айтқанда, мультимедиялық корпусстың репрезентативтілігі, тақырыптарға бөліп тегтеу, корпусстың ыңғайлылығы, техникалық безендірілуі, әдеби тілдің нормасының ережелерін сақтау т.б. Зерттеу нәтижелері қазақ тілінің мультимедиялық корпусының теориялық тұжырымдарын жасап, ішкорпусстарға бөлуге септігін тигізеді.

Кілтті сөздер: корпусстық лингвистика, корпус, мультимедия, ішкорпус, мәтін.

Кіріспе

Қоғам үнемі өзгеріп, даму үстінде болғандықтан, ғылым саласы да сол қоғамның талаптары мен сұраныстарына сай жаңа деңгейге шығады. Соңғы жылдарда корпусстық лингвистика тіл білімінің жаңа бағыты ретінде зерттеліп, өзекті мәселелердің біріне айналды. Соның ішінде мультимедиялық корпусстар және оларды жасау мәселесі зерттеу нысаны ретінде жиі қарастырылуда.

Зерттеу жұмысының мақсаты – мультимедиялық корпус жасау барысында туындайтын проблемаларды анықтау. Қазіргі уақытта әлемде ұлттық корпусстың ішкорпусы ретінде түзілген немесе жеке мультимедиялық корпусстар бар және жасалу үстінде. Алайда отандық мультимедиялық корпусстар жайлы ақпарат зерттеу жұмысына қажетті мәлімет жинау, әдебиеттерге шолу жасау барысында кездеспеді. Сол себепті қазақ тілінің мультимедиялық корпусын жасаудың теориясы мен тәжірибесін қалыптастыру өзекті екені айқындалады.

Мультимедиялық корпус түсінігіне тоқталмас бұрын жалпы корпус дегеннің не екенін анықтап алайық. М. Вейсер берген анықтама бойынша, корпус – мақсаты мен көлеміне байланысты белгілі бір жобалау критерийлер жиынтығына негізделген лингвистикалық анализ жасауда қолданылатын ауызша немесе жазбаша мәтіндердің топтамасы [1, 13]. Корпустық лингвистиканы зерттеген отандық ғалымдар А. Қ. Жұбанов пен А. Ә. Жаңабекова корпуссты кез келген тілдегі электронды форматтағы, бір орталық арқылы басқарылатын автоматтандырылған лингвистикалық мәтіндер жинағы деп анықтайды [2, 13].

Бір сөзбен айтқанда, корпус белгілі бір мақсатқа сай жинақталған цифрланған лингвистикалық мәтіндердің қоры деп сипатталады. Ал құрылу мақсатына қарай корпусстар бірқатар белгілері бойынша ерекшеленуі мүмкін. Г. Кеннеди электронды корпусстардың бірнеше түрін ажыратады: жалпы корпусстар, арнайы корпусстар (бұл корпус өз ішінде бірнеше түрге бөлінеді), жазба тіл корпустары, ауызекі тіл корпустары, синхронды корпусстар, диахронды корпусстар [3, 19]. Соның ішінде технологиялардың дамуына қатысты тілдік зерттеулер мен білім беру саласында тиімді түрде материалдарын пайдалануға мол мүмкіндік беретін лингвистикалық корпусстың бір түрі – мультимедиялық корпус.

«Цифрланған мәтіндерден тұратын тіл корпусынан мультимедиялық корпусстың артықшылығы неде?» деген орынды сұрақ туындауы мүмкін. Алдымен «мультимедиялық корпус» деген терминге түсініктеме беріп көрейік. С. А. Маник өз зерттеуінде В. П. Захаров және С. Ю. Богданованың еңбегіне сүйеніп, «мультимедиялық корпус – бұл айтылу сәтінде «ендірілген» ауызша сөйлеуді үйренуге арналған электрондық ресурс» деп анықтама

береді [4, 19]. Өз кезегінде Дженс Оллвуд «мультимодальді корпус» терминін қолдана отырып, оны бір модалділікке ғана негізделмейтін тілдік және коммуникативтік материалдардың цифрланған топтамасы болып табылады деп есептейді. Ал тар мағынада бұл түсінік аудиовизуалды материалдың транскрипция, аннотация немесе кодтаумен қатар жүруін меңзейді [5, 3].

Демек, мультимедиялық корпуста берілетін мәтіндер жай ғана теріліп, электронды форматта берілген материал емес, ауызша сөйлеудің айтылу жағдайын ескере, оған ене отырып қабылданатын мәтін. Бұл корпус түрі мәтінге қоса оның дыбыстық және бейнеқатарының жүретіндігімен ерекшеленеді. Бір сөзбен айтқанда, мультимедиялық корпус құрастырудың басты себебі – оның түрлі зерттеулерді толығырақ жүргізуге мүмкіндік беретін материалды қамтамасыз ететіндігінде және коммуникацияның құрамдас бөліктерін (ым-ишара мен сөйлеу) бөлмей, бірлестікте қарастыруында.

Қазақ тілінің ұлттық корпустары мыналар: qazcorpus.kz, qazcorpora.kz, <http://web-corpora.net>. Осылардың ішінде qazcorpora.kz сайтында (<https://test.qazcorpora.kz/audio>) қазақ тілін дамытуға үлес қосқысы келетіндерге арнап, аудио жасайтын, дыбыстық жазу батырмасын жасап қойған. Ондағы сөйлемдердің саны бізге беймәлім, бірақ әр түрлі күрделі, жай, бірыңғай т.б. сөйлемдерден тұрады. Ұлттық корпус жасауда мәтіндер мен сөздерді енгізу қаншалықты ұлан-ғайыр жұмыс болса, мультимедиялық корпус жасау да одан бірнеше есе ауыр болмаса, кем болмайды. Сондықтан мұндай корпус жасауға лингвистер ғана емес, еріктілердің де үлес қосуы – құптарлық.

Материалдар мен әдістер

Зерттеудің ғылыми негіздерін қарастыру барысында және әдебиеттерге шолу жұмысы кезінде корпустық лингвистикаға, оның ішінде мультимедиялық корпус жасауға қатысты ғылыми еңбектер қарастырылды. Мультимедиялық корпус жасаудың проблемаларын зерттеп, қандай ғылыми тұжырымдамалар жасалғанын, анықтау, жинақтау және саралау әдістерімен бірге теориялық деңгейдің талдау және салыстыру әдістері қолданылды.

Нәтижелер мен талқылау

Мультимедиялық корпусты жасауда туындайтын қиындықтар жайлы айтпастан бұрын НКРЯ (Орыс тілінің ұлттық корпусы) құрамына енетін мультимедиялық корпустың (МУРКО) негізінде корпустың бұл түрі жайлы түсінік берген жөн. Қазіргі уақытта аталған корпус үш негізгі өлшем бойынша мәлімет іздеуге мүмкіндік береді: айтылатын мәтін бойынша, сөйлеу әрекетінің типіне байланысты (келісу, алғыс білдіру, құттықтау және т.б.), сонымен қатар ым-ишара арқылы (пайдаланушы интерфейсіні жасауда ыңғайлы болу үшін корпусты құрастырушылар ым-ишара деп ауызша сөзбен қатар жүретін вербалды емес сүйемелдеу рөліндегі кез келген қимылды түсінеді) [6].

Корпус көлемі шамамен 4,5 млн сөзқолданысқа жақын. МУРКО құрамы бес ішкорпусқа бөлінеді:

1 Кино тілі. Бұл бөлім 1930–2000 жылдардағы кеңес және ресейлік кинофильмдерді қамтиды. Жалпы көлемі 3,4 млн сөзқолданысты құрайды.

2 Бұқаралық ауызекі тіл. Аталған ішкорпус өте қарқынды даму үстінде. Бұл бөлімде толыққанды зерттелгені - ғылыми ауызекі тіл: конференцияларда жасалған баяндамалар мен талқылаулар, теледидардағы танымал дәрістер, радиодағы ток-шоулар мен бағдарламалар т.б.

3 Күнделікті қолданыстағы ауызекі тіл ішкорпусы диалогтар мен шағын диалогтарды, отбасылық және достар арасындағы, телефон әңгімелерін және т.б. қамтиды.

4 Театр саласының ауызекі тілі сахналанған және радиодан берілетін театр қойылымдары арқылы көрініс табады.

5 Авторлық және көркем оқу бөлімдері дыбысталған жазба тілді ұсынады. Бұл дыбыстық мәтіннің фонетикалық ерекшеліктерін, орфоэпия және акцентология, интерпретациясын зерттеуде аса маңызды.

Аталған корпуста қамтылған материал 10–30 секундтық кішігірім бөліктерге (клип) бөлінген аудио және бейнефайл түрінде беріледі, әрқайсысында мәтіндік декодтау бар. Клип және мәтіннің бірлестігі аяқталған мағыналы коммуникативті бөлік болып табылады. Бөлік корпустың стандарттарына сай белгіленеді және сайттан іздеуді жүзеге асыруға мүмкіндік беретін метамәтіндік, акцентологиялық, морфологиялық, семантикалық аннотациясы жасалынады [7].

Көрсетілген мультимедиялық корпустың қысқаша сипаттамасына назар аударар отырып, лингвистиканың корпустық бағыты, әсіресе мультимедиялық корпустарды зерттеу саласы әлеуеті күшті және ізденіске көп мүмкіндік беретініне көз жеткізеді.

Мультимедиялық корпустардың кеңінен қолданылып, тіл білімі саласында бірқатар мәселелерді шешуге септігін тигізетіні жайлы Кэтрин Акерли мен Франческа Коччеттаның зерттеуінен де тұжырымдауға болады. Авторлардың айтуынша, корпустарды қолдану тілді меңгеруге, сөздіктер мен грамматикаларды құрастыруға зор үлес қосады. Оны Акерли мен Коччеттаның зерттеуінен ағылшын тілінің мультимедиялық корпусы негізінде тілді меңгеру үйренушілерге ауызекі ағылшын тілінің шынайы қолданысын көруге мүмкіндік беретініне көз жеткізуге болады. Жай мәтінге қарағанда бейнеклиптерді көру тілді қолданғанда кез келген жағдайды талдау жасауға әсерін тигізетін лингвистикалық, жағдаяттық, әлеуметтік, психологиялық және прагматикалық факторларға ауқымдырақ қолжетімділік береді [8].

Мультимодальді корпус тек жазба тілді қамтумен шектелмей, контекстуалды және мәдени сипаттағы мәліметті қосуға да мүмкіндік береді. Демек, бұл мультимодальді корпус тар мәдени және тілдік үрдістерді тұтастықта құжаттауда таптырмас құрал болып табылады [5].

Мультимедиялық корпус мәселесінің балалар тілін зерттеуде де маңызы бар. Балалар тілінің мультимедиялық корпусы жайлы еңбегінде Дэн Сянцзюнь мен Вирджиния Ип мультимедиялық корпус зерттеушілерге баланың лингвистикалық және лингвистикалық емес өнімін, сонымен қатар қажетті контекстуалды мәліметті белгілеп отыруға мүмкіндік беретінін пайымдайды [9]. Тілді ұрпақтан-ұрпаққа таза, бұзылмаған күйінде сақтап, мұра етіп қалдыру тілші мамандардың еншісінде. Ана тілі дұрыс шығып, негізі қаланса, қосалқы тілдер зиянын тигізе бермейді. Сол себепті бала тілінің жағдайына көп мән беру керек. Осыған сәйкес мультимедиялық корпус негізінде зерттеулер жасалып, балалар тілінің нормалануы, осы саладағы шешілмеген проблемалар мен сұрақтар айқындалады.

Келтірілген фактілер мультимедиялық корпус құрастыру мәселесінің маңыздылығын көрсетіп, жұмыс барысында туындайтын проблемаларды анықтауды талап етеді.

Корпус құрастыруда қандай проблемалар туындауы мүмкін және олар неге байланысты? Біріншіден, кез келген корпусстың репрезентативті болуы. Корпусстың репрезентативтілігі деп үлкен ауқымды ақпараттың көлемі бойынша едәуір кіші деректер корпусымен бірдей көрініс тауып, бейімделуін немесе бірігіп кетуі ретінде қарастыруға болады [10, 116]. Нақтырақ айтқанда, корпус аясында көп мәліметті қысқа әрі қолжетімді етіп ұйымдастыру үшін мультимедиялық корпусстың құрылымын (архитектурасын) жан-жақты ойластырып, сауатты түрде құру керек.

Корпус құрастыру жұмысы үлкен еңбек пен уақыт алатыны белгілі. Корпусты дыбыстаудың өзі біршама уақытты талап етеді. Сонымен қатар тілдік материалмен жұмыс істеу үшін тиімді бағдарламалық қамтамасыз ету, аннотацияларды стандарттау, корпусстың қолжетімділігі мен көпсалалығын қамтамасыз ету және қойылған мақсатқа қол жеткізу үшін қаржыландырудың жеткілікті болуы да аталған жұмысты атқаруда туындайтын проблемалар қатарына жатады [11]. Кристоф Парисс және Алия Моргенштернның көзқарастары да басқа зерттеушілердің тұжырымына ұқсас, басты нәрсе – бұл зерттеушінің корпус құрастыру барысында атқаратын жұмысқа (аудиожазба жасау, транскрипция, байланыстыру, өңдеу) жұмсалатын еңбегі мен уақыты [12].

Қажырлы еңбек пен уақыт ресурстарының көп жұмсалыу мәселесі корпус жасау барысында туындайтын жалғыз қиындық емес. Барлық заманауи корпус тар оқуға ыңғайлы болуы тиіс, ал бұл белгілерге берілген

аннотациялар тек лингвистикалық теорияға емес, сонымен қатар оларға қатысты тегтерді (мәліметтер базасындағы әрбір тақырыптың және ауқымды тақырыптың сандық идентификаторы) [13, 47] бірізді және жүйелі түрде қолдануға негізделуі керек [14].

Корпус жақсы құжатталып (яғни нақты және қатысы бар метадеректерден құралуы), тиісті регистрлер мен әлеуметтік – лингвистикалық деректерді қамтуы тиіс. Аталғанның бәрі арнайы технологияларды (мысалы, ELAN), стандарттар мен хаттамаларды (мысалы, IMDI метадеректер дескрипторлары), сондай-ақ ашық және келісілген грамматикалық тегтерді (мысалы, грамматикалық кластардың белгілері) талап етеді [14, 82].

Мультимедиялық корпус тек ауызекі тілге қатысты мәселерді қарастырумен шектелмейді. Корпустық лингвистиканың дамуы, ең алдымен, мультимедиялық корпустарды жаппай құру вербалды емес жағдайларды зерттеуге мүмкіндік береді, яғни ым-ишара, дене қимылы іспетті ауызша сөйлеу компоненттері жүйелі негізде қарастырылады [15]. Адам бейвербалды амалдарсыз сөйлеуі мүмкін емес. Дененің қимылынан бөлек, бет-ауыздың жыбырлауы, қасы-көздің құбылуы, барлығы сөйлеу әрекеті барысында бірге ілесіп отырады. «Мультимедиялық корпус жасау кезінде сөз бен қимылды біріктіріп, бір жүйеге салуға бола ма?» деген мәселелер зерттеу жұмыстарына арқау болуда. Бұл мақала аясында қарастырылмаған, коммуникацияның вербалды емес құралдарын (ым-ишара, дене қимылдары) келесі зерттеулердің тақырыбы ретінде алу көзделіп отыр.

Қорытынды

Қорыта айтқанда, мультимедиялық корпус жасау мәселесі бүгінгі күні тіл білімі үшін аса маңызды. Корпустың бұл түрі лингвистикалық зерттеулер жүргізуде, сөздіктер құрастыруда, тілді үйретуде т.б. тиімді пайдаланылады. Яғни, тіл тек цифрланған мәтіндегі мәлімет арқылы емес, ауызекі сөйлеу тілінің түрлі жағдайдағы қолданысын ескере, оған ене отырып қабылданады, зерттеледі. Мультимедиялық корпустағы материалдың дыбысталуы және бейне арқылы толықтырылуы оның аутентикалық құндылығын арттырады және нақты қолданысын айқындайды, сондай-ақ вербалды қатынас құралдарын вербалды емес құралдарынан ажыратпай қарастыруға мүмкіндік береді.

Мультимедиялық корпустың таза, табиғи мәліметтері арқылы көстілді, көптілді сөйлеушінің қоғамдағы әлеуметтік маңызы мен шектеулерін зерттеуге мүмкіндік береді. Эмпирикалық зерттеу жасағанда табиғи корпустың мәліметтері негізгі жиілік деректер ретінде бақыланатын мәліметтермен үйлестіріле қолданылады.

Алайда кез келген зерттеу барысында қиындықтар мен проблемалар туындайды. Мультимедиялық корпус жасау барысында да бірқатар

проблемалар кездесетінін анықтадық. Соның негізінде бұл проблемаларды топтастырып отырмыз:

- 1 Мультимедиялық корпусның репрезентативтілігін қамтамасыз ету.
 - 2 Корпуста қамтылатын материалды үйлестіру және тег жасауға ыңғайлап тақырыптарға бөлу.
 - 3 Корпусның оқуға ыңғайлылығы және қолжетімділігі.
 - 4 Корпусның құжатталуы мен техникалық безендірілуі.
 - 5 Корпус жасауда әдеби тіл нормасын сақтау.
- Көрсетілген проблемаларды ескеріп, есепке ала отырып, мультимедиялық корпус жасау үрдісін жеңілдетуге және жүйелендіруге болады деген қорытындыға келдік.

Пайдаланылған деректер тізімі

1 **Weisser, M.** Practical Corpus Linguistics : An Introduction to Corpus-Based Language Analysis [Text]. – Wiley Blackwell : 2016. – 150 p.

2 **Жұбанов, А. Қ., Жаңабекова, А. Ә.** Корпустық лингвистика [Мәтін]. – Алматы : «Қазақ тілі» баспасы, 2016. – 336 б.

3 **Kennedy, G.** An Introduction to Corpus Linguistics. Published / Book Reviews : An Introduction to Corpus Linguistics (nus.edu.sg) – [Electronic resource]

4 **Маник, С. А.** Современная корпусная лексикография: особенности и перспективы развития [Текст] // Вестник Ивановского государственного университета. Серия : Гуманитарные науки. – 2016. – № 3. – С. 17-20.

5 **Allwood, J.** Multimodal Corpora. Corpus Linguistics [Text]. – An International Handbook, Mouton de Gruyter. – 2008. P. 207-225.

6 **Соболева, О. В.** Обучающий потенциал мультимедийного корпуса в составе национального корпуса русского языка [Текст] // Новое и традиционное в переводоведении и преподавании русского языка как иностранного. Сборник статей. Панъевропейский университет – «АПЕИРОН», Баян-Лука, – 2022. – С. 77–84.

7 **Савчук, С. О., Махова, А. А.** Мультимедийный поэтический корпус: принципы создания и особенности разметки [Текст] // «Слово и жест» Научная конференция, посвященная памяти Е. А. Гришиной («Гришинские чтения»). М. : Институт русского языка им. В. В. Виноградова РАН, 2020. – 50 с.

8 **Ackerley, K., Coccetta, F.** Enriching language learning through a multimedia corpus [Text]. // ReCALL. – 2007. – № 3. – Т. 19 – P. 351–370.

9 **Xiangjun, D., Yip, V. A.** multimedia corpus of child Mandarin: The Tong corpus [Text] // Journal of Chinese Linguistics. – 2018. – № 1. – Т. 46. – P. 69–92.

10 **Климчук, А. И., Рыков, В. В., Склярова, Е. В.** Когнитивные технологии и мультимедийные корпуса [Текст] // СРТ 2018 Труды Международной научной конференции по физико-технической информатике. – Москва-Протвино, 2018. – С. 115–117.

11 **Симакова, М. Ю.** Лингвистические проблемы создания речевого корпуса. Русские речевые корпуса. [Текст] // Международная научно-техническая конференция, приуроченная к 50-летию МРТИ-БГУИР (Минск, 18–19 марта 2014 года) : Материалы конф. В 2 ч. Ч. 1 -Минск, 2014. – С. 310–311.

12 **Parisse, C., Morgenstern, A. A.** multi-software integration platform and support for multimedia transcripts of language [Text] // LREC 2010: Workshop on Multimodal Corpora : Advances in Capturing, Coding and Analyzing Multimodality. – 2010. – P. 106–110

13 **Земичева, С. С., Иванцова, Е. В.** Тематическая разметка диалектного корпуса: опыт Томских диалектологов [Текст] // Вестник Томского государственного университета. Филология. – 2020. – № 67. – С. 45–61.

14 **Johnston, T.** Corpus linguistics and signed languages: no lemmata, no corpus [Text] // European Language Resources Association (ELRA), 2008. – P. 82–87.

15 **Гришина, Е. А.** Указания рукой как система (по данным Мультимедийного русского корпуса) [Текст] // Вопросы языкознания. – Москва, Российская академия наук, «Наука», – 2012. – № 3. – С. 3–50.

References

1 **Weisser, M.** Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis [Text]. – Wiley Blackwell, 2016. – 150 p.

2 **Jubanov, A. Q. Janabekova, A. Á.** Korpustyq lingvistika [Corpus linguistics] [Text]. – Almaty : «Qazaq tili» baspasy, 2016. – 336 p.

3 **Kennedy, G.** An Introduction to Corpus Linguistics. Published / Book Reviews: An Introduction to Corpus Linguistics (nus.edu.sg) – [Electronic resource]

4 **Manik, S. A.** Sovremennaja korpýsnaia leksikografija: osobennosti i perspektivy razvítia [Modern corpus lexicography: features and prospects of development] [Text] // Vestnik Ivanovskogo gosýdarstvennogo ýniversiteta. Seria : Gýmanitarnye náýki. – 2016. – № 3. – P. 17–20.

5. **Allwood, J.** Multimodal Corpora. Corpus Linguistics [Text] – An International Handbook, Mouton de Gruyter. – 2008. P. 207–225.

6 **Soboleva, O. V.** Obýchaiýshi potentsial mýltimedijnogo korpýsa v sostave nasionalnogo korpýsa rýsskogo iazyka [The educational potential of the multimedia

corpus as part of the national corpus of the Russian language] [Text] // *Novoe i tradisionnoe v perevodovedenii i prepodavanii rýsskogo iazyka kak inostrannogo. Sbornik statei Panevropeiskii úniversitet – «APEIRON», Baná-Lýka, – 2022. – С. 77–84.*

7 **Savchuk, S. O., Mahova, A. A.** Multimedyny i poeticheskii korpus : printsipy sozdaniia i osobennosti razmetki [Multimedia poetic corpus : principles of creation and features of markup] [Text] // «Slovo i jest» Naýchnaia konferentsia, posveshennaia památi E. A. Grishinoi («Grishinskie chtenia»). – M. : Instítýt rýsskogo iazyka im. V. V. Vinogradova RAN, 2020. – 50 p.

8 **Ackerley, K., Coccetta, F.** Enriching language learning through a multimedia corpus [Text]. // *ReCALL. – 2007. – № 3. – Т. 19 – P. 351–370.*

9 **Xiangjun, D., Yip, V. A.** multimedia corpus of child Mandarin: The Tong corpus [Text] // *Journal of Chinese Linguistics. – 2018. – № 1. – Т. 46. – P. 69–92.*

10 **Klimchuk, A. I., Rykov, V. V., Skharova, E. V.** Kognitivnyie tehnologii i multimedynye korpusa [Cognitive technologies and multimedia enclosures] [Text] // *CPT 2018 Trýdy Mejdýnarodnoi naýchnoi konferentsii po fiziko-tehnicheskoi informatike. – Moskva-Protvino, 2018. – P. 115–117*

11 **Smakova, M. Iy.** Lingvisticheskie problemy sozdaniia rechevogo korpusa. Russkie rechevye korpusa [Linguistic problems of creating a speech corpus. Russian speech corps] [Text] // *Mejdýnarodnaia naýchno-tehnicheskaiia konferentsia, priýrochennaia k 50-letíy MRTI-BGÝIR (Minsk, 18–19 marta 2014 goda) : Materialy konf. V 2 ch. Ch. 1 -Minsk, 2014. – P. 310–311.*

12 **Parisse, C., Morgenstern, A. A.** multi-software integration platform and support for multimedia transcripts of language [Text] // *LREC 2010: Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. – 2010. – P. 106–110.*

13 **Zemicheva, S. S., Ivantsova, E. V.** Tematicheskaiia razmetka dialektного korpusa: opyt Tomskih dialektologov [Thematic markup of the dialect corpus: the experience of Tomsk dialectologists] [Text] // *Vestnik Tomskogo gosýdarstvennogo úniversiteta. Filologia. – 2020. – № 67. – P. 45–61.*

14 **Johnston, T.** Corpus linguistics and signed languages : no lemmata, no corpus [Text] // *European Language Resources Association (ELRA), 2008. – P. 82–87.*

15 **Grishina, E. A.** Ukazaniia rukoi kak sistema (po dannym Multimedynogo russkogo korpusa) [Hand instructions as a system (according to the Multimedia Russian Corpus)] [Text] // *Voprosy iazykoznaniiia. – Moskva, Rossiiskaia akademiia naýk, «Naýka», – 2012. – № 3. – P. 3–50.*

Материал 13.03.23 баспаға түсті.

*Г. О. Сакиева¹, Л. Ә. Ибраймова², Б. Молдағали³

^{1,2}Қазақский национальный женский педагогический университет,
Республика Казахстан, г. Алматы;

³Қазақский национальный педагогический университет имени Абая,
Республика Казахстан, г. Алматы.

Материал поступил в редакцию 13.03.23.

ПРОБЛЕМЫ СОЗДАНИЯ МУЛЬТИМЕДИЙНОГО КОРПУСА

Создание мультимедийного корпуса казахского языка – это новая область, требующая проведения большого объема исследований. Разговорный язык до сих пор рассматривался только на уровне сравнения, анализа текста. Мультимедийный корпус может стать очень эффективным инструментом в освоении казахского языка как второго. Может внести большой вклад в выявление отклонений и несоответствий в норме разговорного и литературного языка. В данной статье выявлены проблемы, возникающие при создании мультимедийного корпуса. Выявление проблем значительно облегчит работу по созданию корпуса и поможет заранее определить и преодолеть сложности. Для этого было дано краткое определение термина «корпус», в том числе определение понятия «мультимедийный корпус». На примере Национального корпуса казахского языка и Мультимедийного русского корпуса (МУРКО), входящего в состав НКРЯ (Национального корпуса русского языка), было сделано краткое описание. Опираясь на выводы ученых, изучавших эту область, была определена важность и необходимость мультимедийных корпусов. Также были рассмотрены сложности, возникающие при создании корпуса. Они были разделены на несколько типов. В частности, репрезентативность мультимедийного корпуса, разбивка на теги, доступность, техническое оформление корпуса, соблюдение правил норм литературного языка и т.д. Результаты исследования способствуют созданию теоретических концепций мультимедийного корпуса казахского языка и его разделению на подкорпусы.

Ключевые слова: корпусная лингвистика, корпус, мультимедия, подкорпус, текст.

*G. O. Sakieva¹, L. A. Ibraimova², B. Moldagali³

^{1,2}Kazakh National Women's Pedagogical University,
Republic of Kazakhstan, Almaty;

³Abai Kazakh National Pedagogical University,
Republic of Kazakhstan, Almaty.

Material received on 13.03.23.

PROBLEMS OF CREATING A MULTIMEDIA CORPUS

The creation of a multimedia corpus of the Kazakh language is a new area that requires a large amount of research. The spoken language has so far been considered only at the level of comparison, text analysis. Multimedia corpus can become a very effective tool in mastering the Kazakh language as a second language. It can make a great contribution to the identification of deviations and inconsistencies in the norm of spoken and literary language. This article identifies the problems that arise when creating a multimedia corpus. The identification of problems will greatly facilitate the work on the creation of the corpus and will help to identify and overcome difficulties in advance. For this purpose, a brief definition of the term «corpus» was given, including the definition of the term «multimedia corpus». National Corpus of the Kazakh language and Multimedia Russian Corpus (MURKO), which is part of the NKRYA (National Corpus of the Russian Language), were used as an example to make a brief description. Based on the conclusions of scientists who have studied this area, the importance and necessity of multimedia corpora was determined. The difficulties arising during the creation of the corpora were also considered. They were divided into several types. In particular, the representativeness of the multimedia corpus, division into tags, accessibility, technical design of the corpus, compliance with the rules of the norms of the literary language, etc. The results of the research contribute to the creation of theoretical concepts of the multimedia corpus of the Kazakh language and its division into subcorpora.

Keywords: corpus linguistics, corpus, multimedia, subcorpus, text.

Теруге 13.03.2023 ж. жіберілді. Басуға 31.03.2023 ж. қол қойылды.

Электронды баспа

3,41 МБ RAM

Шартты баспа табағы 25,38. Таралымы 300 дана. Бағасы келісім бойынша.

Компьютерде беттеген: А. К. Темиргалинова

Корректорлар: А. Р. Омарова, Д. А. Кожас

Тапсырыс № 4037

Сдано в набор 13.03.2023 г. Подписано в печать 31.03.2023 г.

Электронное издание

3,41 МБ RAM

Усл. печ. л. 25,38. Тираж 300 экз. Цена договорная.

Компьютерная верстка: А. К. Темиргалинова

Корректоры: А. Р. Омарова, Д. А. Кожас

Заказ № 4037

«Toraighyrov University» баспасынан басылып шығарылған

Торайғыров университеті

140008, Павлодар қ., Ломов к., 64, 137 каб.

«Toraighyrov University» баспасы

Торайғыров университеті

140008, Павлодар қ., Ломов к., 64, 137 каб.

67-36-69

e-mail: kereku@tou.edu.kz

www.vestnik.tou.edu.kz