МРНТИ 16.31.21

А.Қ. Жұбанов

А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің бас ғылыми қызметкері,филология ғылымдарының докторы, профессор. Алматы қаласы, Қазақстан

ЦИФРЛЫҚ ҚАЗАҚСТАН ҮШІН ҚАЗАҚ ТІЛІНІҢ СӨЙЛЕУ КОРПУСЫН ҚҰРУ ӨЗЕКТІ МӘСЕЛЕ

Аннотация: Мақалада соңғы онжылдықтағы әлемдік тілдердің «Сөйлеу тілі корпустарын құру» мәселесіне қысқаша шолу жасалып, қолданылып келе жатқан әдістердің корпустық модельдеу мен сөйлеу тілін автоматты синтездеу салаларына қарай ауыса бастағандығы жайлы сөз болады. Бұл жағдайдың сөйлеу тілінің просодикалық сипаттамасын, оның эмоционалды мазмұнын модельдеу үшін аса маңызды екені айтылған

Сонымен бірге, мақалада, болашақта құрастырылатын қазақ тілінің сөйлеу корпусын тек ғылыми зерттеу мақсатында пайдаланумен бірге дискретті сөйлеу тілінің бірліктерін автоматты түрде тану жүйесін құру мәселесінің де шешімі табылатыны жайлы сөз болады.

Мақалада «Сөйлеу тілі корпусын» құру мәселесінің әлемдік деңгейдегі тәжірибесі мен маңыздылығы қысқаша баяндалады. Аталған корпустың ғылыми зерттеу мақсатында пайдаланылуымен бірге дискретті сөйлеу тілінің бірліктерін автоматты түрде тану жүйесін құру үшін де маңыздылығы қысқаша сөз болады.

Тірек сөздер: Корпустық лингвистика, компьютерлік бағдарлама, корпустар базасы, корпустық модельдеу, сөйлеу тілі корпусы, сөйлеу қоры, сөйлеу сигналы, тілдік ресурс, акустикалық вариативтілік, жасанды интеллект, автоматты тану, дыбыстық сигнал, фрейм, сөйлеу тілін синтездеу, дискретті сөйлеу, үзіліссіз сөйлеу, фонемалық/фонетикалық/просодикалық транскрипция, автоматты синтездеу, эмоционалды мазмұн, сөйлеу технологиясы.

А.К. Жұбанов

главный научный сотрудник Института языкознания имени А. Байтурсынова, док.филол.н., профессор, Алматы, Казахстан

ДЛЯ ЦИФРОВОГО КАЗАХСТАНА СОЗДАНИЕ КОРПУСА УСТНОЙ РЕЧИ КАЗАХСКОГО ЯЗЫКА ЯВЛЯЕТСЯ АКТУАЛЬНОЙ ПРОБЛЕМОЙ

Аннотация. В статье дается краткий обзор проблемы «создания корпуса устной речи» в мировых языках последнего десятилетия, а также обсуждается переход существующих методов в область моделирования корпусов и автоматического синтеза устной речи. Было отмечено, что данная ситуация очень важна для моделирования просодической характеристики устной речи, ее эмоционального содержания. Кроме

того, в статье рассматривается решение проблемы создания системы автоматического распознавания дискретных речевых единиц, а также использование корпуса казахского языка в исследовательских целях в будущем. В статье обобщается опыт мирового уровня и важность создания «корпуса устной речи». Помимо использования этого корпуса в научно-исследовательских целях, важно также создать систему автоматического распознавания дискретных речевых единиц.

Ключевые слова: Корпусная лингвистика, компьютерная программа, корпусная база, корпусное моделирование, корпус устной речи, речевая база, речевой сигнал, языковой ресурс, акустическая вариативность, искуственный интеллект, автоматическое распознавание, акустический сигнал, фрейм, синтез речи, дискретная речь, непрерывная речь, фонемная/ фонетическая/просодическая транскрипция, автоматическое синтезирование, эмоциональное содержание, технология речи.

A.K. Zhubanov

Chief Researcher of the A. Baitursynuly Institute of the Linguistics,
Doctor of Philology, Professor, Almaty, Kazakhstan

FOR DIGITAL KAZAKHSTAN CREATION OF THE SPEAKER CORPS OF THE KAZAKH LANGUAGE IS A TOPICAL PROBLEM

Annotation. The article at the world level gives a brief overview of the problems of creation and the importance of using the corpus of oral speech in the Kazakh language. In addition, the article talks about the importance of using the corpus of oral speech in the Kazakh language both for the purposes of scientific research and for the purpose of creating a system for automatic recognition of units of discrete oral speech in the Kazakh language.

Keywords: Corpus linguistics, corpus of oral speech, computer program, speech base, corpus base, corpus modeling, speech signal, language resource, acoustic variability, artificial intelligence, automatic recognition, frame, speech synthesis, discrete speech, continuous speech, phonemic / phonetic / prosodic transcription, automatic synthesis, emotional content, speech technology.

Қазақ тілін зерттеуде корпустық лингвистика саласына көп мән берілуі және оның әлемдік дәрежедегі теориялық және практикалық жақтарын зерттеу қажеттігі туындауда.

Әлем бойынша мәтін корпустарын құрастыру мен олардың қызметіне қатысты жалпы және нақты мәселелерге арналған мақалалар жарық көрген ғылыми журналдардың арнайы басылымдары да шыға бастады [1].

Бірақ әлде де қазақ тіл білімі үшін корпустық лингвистикаға қатысты көптеген мәселелер арнайы зерттеуді қажет ететіні белгілі. Оған жататындар: корпустық лингвистика мен оның негізгі ұғымдарының анықтамалары, корпустық лингвистиканың тіл білімі құрылымында алатын орны, әдіс-тәсілдері және т.б. Сонымен бірге жаңа бағыттың теориялық негізін ұғыну мәселесі корпустарды нақты зерттеулерде пайдалануға қарағанда белгілі дәрежеде қалыс қалып келе жатқаны да байқалады.

Корпустық лингвистика пәнін осы саланың мамандары тілдік корпустарды құру мен оны пайдалану жағдайын зерттейтін тіл білімінің бір саласы ретінде ғана қарас-

тырып келді. Кейбір ғалымдар ол пәннің түсінігін тар шеңберде қарастырып, оны тек компьютерлік лингвистика саласының аясында ғана түсіндіреді: «Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий» – дейді [2].

Ал компьютерлік лингвистика ұғымын, әдетте, компьютерлік құралдарды пайдаланудың кең мүмкіндігі ретінде түсіндіруге болатыны белгілі. Бұл жердегі «компьютерлік құралдар» деп отырғанымыз – компьютерлік бағдарламалар, тілдік деректерді өңдеу мен компьютерлік технологияны орынды ұйымдастыру жұмыстары және т.б. [3].

Ал корпустық лингвистика компьютерлерді тек «құрал» ретінде пайдалананады. Міне, сондықтан да корпустық лингвистика өзіне жүктелген міндетті ондай құралсыз атқара да алмас еді. Бірақ компьютер мұндай рөлді қазіргі білім саласының барлық түрлерінде де атқаратынын ескерсек, онда олардың бәрін бірдей компьютерлік лингвистика саласына жатқыза беруге де болмайды.

Жоғарыда сөз болған корпустық лингвистика пәнінің теориялық және тәжірибелік жақтары қазақ тілі мәтіндері бойынша компьютерлік корпустар базасын құру жағдайында да ескерілуі қажет. Корпустық лингвистика қазақ тіл білімінің ерекше саласы ретінде қалыптасатын болса, қазақ тілі мамандарына көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тауып алуға және оларға тиісті деген өңдеулер жүргізуге мүмкіндік туындатады. Осының бәрі қазақ тіліне қатысты зерттеулердің шынайылыққа (ақиқаттыққа) жетудің эмпирикалық тәсілдеріне жаңаша көзқараспен қарауға және ғылыми айналым аясына аса маңызды тілдік материалдарды енгізуге жағдай жасайды.

Қазіргі кезде әлемдік корпустық лингвистиканың даму сипаты – ұлттық толық мәтіндерді арнайы зерттеу нысаны етіп алу. Сондықтан автоматтанған қазақ тіліндегі мәтіндер корпусының компьютерлік базасы (теориялық және практикалық жағынан қарастырғанда) жақын болашақта жүзеге асатын «Қазақ тілінің ұлттық корпусының» аса маңызды бастамасы болары сөзсіз. Мұндай зерт-теулердің нәтижелері қазақ мәтіндерінің стильдік, құрылымдық, мағыналық, функционалдық және т.б. сипаттарын анықтауда да өзекті мәселелердің бірі болып саналады.

Енді қазақ тіл білімінде әлі қарастырыла қоймаған «**Сөйлеу тілі корпусы**» жайлы осы мақалада қысқаша сөз етпекпіз.

Дыбыстама сөйлеу корпустарын деректердің сөйлеу қоры деп те атайды және оны тілдік ресурстардың маңызды бір түрі ретінде де санайды. Корпустың құрамына компьютерлік бағдарламаларды да қоса есептеу жиі кездеседі, оның себебі ондай бағдарламалар тілдік, оның ішінде фонетикалық ресурстарды құру, жинау, ұйымдастыру мен басқару әрекеттерін қамтамасыз ететін құрал. Сөйлеу корпустарын құруға қызығушылық бастама болған, негізінен, сөйлеу тілін автомат-ты түрде тануға қатысты жүргізілген зерттеулер аясы деуге болады. Себебі, бұл салада зерттеушілер тілдің дыбыстық бірліктерінің көптеген акустикалық вариативтілігімен жиі кездесіп отырады. Ал мұндай вариативтілік алуан түрлі дереккөздерде кездесетіні белгілі. Мысалы, оларға жататындар – сөйлеушінің немесе сөйлеу материалын жазуға арналған микрофонның сипаттамасына қатысты психофизиологиялық күйіне дейінгі жүйелік контекстік варитивтілік. За-манауи сөйлеу тілін тану жүйелері, әдетте, таспаға жазылып алынған көптеген дикторлардың (100-ден көп) аса үлкен дыбыстама

сөйлеу ауқымдары (массив) арқылы үйретіледі. **Сөйлеу тілі корпусы** дегеніміз – сөйлеу тілінің құрылымдан-ған бөліктерінің жиынтығы. Мұндай мәліметтермен корпус бойынша әрекет ету арнайы жазылған компьютерлік бағдарламалар арқылы қамтамасыз етіледі. Ал **сөйлеу тілінің бөліктерін** базалық бірлік ретіндегі сөйлеу сигналының цифрланған бөлігі деп және ассоцияланған сипаттағы ақпараттың бір түрі ретінде қабылдау қажет.

Қазіргі кезде сөйлеу корпусын көлемді, көп түрлі және ақпаратты иемдену жағынан бай (көп салалы), сонымен бірге құрастыру мен пайдалану жақтарын ұтымды ету іргелі фонетикалық зерттеулер үшін аса өзекті болуда.

Сол сияқты «Жасанды интеллект» саласындағы зерттеулер де қазіргі кезде әлем бойынша көпшілік ғалымдарға зор қызығушылық тудыруда. Атап айтқанда, жасанды интеллект саласы машиналық (компьютерлік) оқытумен тығыз қатынаста. Аталған сала, яғни жасанды интеллект саласы ғылыми тәжірибеде кең қолданыс табуда және тілдік бейнелерді автоматты түрде тануда көптеген мәселелердің шешімін табуда (Pattern Recognition). Тілдік бейнелерді автоматты тану дегеніміз – ол тілдік бейнелерді бірнеше категориялар немесе кластар бойынша топтастырумен айналысатын ғылыми пән. Мәселен, фонетика ғылым саласы бойынша сөйлеу тілін ойдағыдай тану үшін, әдетте, дыбыстық сигналдың фрейм деп аталатын бір-неше миллисекунд аралығындағы бейнесін ғана қарастырады екен.

Болашақта, сапалы түрдегі автоматты тану мен сөйлеу тілін синтездеу әдістері көлемді түрдегі оқыту саласына арналған компьютерлік ойындарға негіз болуы да мүмкін. Мұндай ойындарда оқушылар компьютермен өздері білетін табиғи тілде қарым-қатынас жасап, ал арнайы жазылған компьютерлік бағдарлама оқушылардың дыбыстық айтылымы мен лексика-грамматикалық рәсімделуіне тиісті баға берумен айналысады деп ойлаймыз.

Сөйлеу корпустарын топтастыру жайында. Сөйлеу корпустарын құрастыру мен оларды пайдалануға қатысты жинақталған тәжірибе бірнеше белгілерді бөліп алуға мүмкіндік тудыра отырып, сөйлеу деректер қорын топтастыруға негіз болады және ол деректер жаңа сөйлеу корпусын жобалау кезінде ескеріледі. Енді ондай топтастырудың ең маңызды сипаттамаларына тоқталайық [4].

Сөйлеу корпусын мақсатты пайдалану:

– мамандандырылған, жалпы (репрезентативті), оқыту және иллюстрациялық;

– сөйлеу материалының түрі: дискретті сөйлеу, үзіліссіз сөйлеу және оқу, өздігінен сөйлеу (спонтанная речь), арнайы диалогтар;

– мәтіндік материалдар түрі: сөздер/буындар тізімдері, жеке сөйлемдердің жиынтығы, өзара байланыстағы мәтіндер; көптақырыптық (монотематикалық) немесе көпфункционалды;

Сөйлеу сигналының түрі: зертханалық сөйлеу, кеңсе сөзі, көпшілік алдында сөйлеу, телефон арқылы сөйлесу (әдетті немесе ұялы телефон арқылы, радио, теледидарлық сөйлеу).

– ***дыбыстық сигналмен байланысты ақпарат түрі (аннотациялар):*** орфографиялық жазба, фонемалық/фонетикалық транскрипция, просодикалық транскрипция, сигналдың акустикалық-фонетикалық белгіленімдер: «оқиғалық», сегменттік, просодикалық, лингвистикалық аннотациялар мен пікірлердің басқа түрлерінің болуы, мысалы, жеке ерекшеліктер туралы сөз иесінің сөйлеуі немесе сөйлеу бөліктерінің эмоционалдық бояуы;

– ***тілдің дыбыстық бірліктерініңстатистикалық теңдестіру түрі:*** табиғи, біркелкі, репрезентативті (өкілдік), арнайы статистикалық сызбаға сәйкес;

– сөйлеу корпусының жадына енгізілген ***қосымша сигналдық ақпараттың болуы және олардың түрі:*** дыбыстық сигналмен қатар, қарапайым, мультимодальды және арнайы құрастырылған сөйлеу корпустары.

Әдетте, сөйлеу дерекқорлары көптілділік сипатта болып келеді. Сөйлеу корпустары тек барлық технологиялық маңызды тілдерге (американдық ағылшын, неміс, жапон, қытай және т.б.) ғана емес, сондай-ақ Еуропалық Одақтың ресми тілдерінің көпшілігінде: ағылшын, голланд, дат, швед, неміс, француз, итальян, испан және басқа да бірнеше тілдік корпустар үшін құрастырылған деуге болады.

Сөйлеу корпустары үшін жазылған Copernicus ELRA корпустық бағдарламасының жүзеге асырылуының нәтижесінде Шығыс Еуропа тілдері (поляк, болгар, эстон, румын және венгер) үшін сөйлеу корпустары пайдаланыла бастады. Интернеттегі Еуропа Қауымдастығының веб-сайтынан орыс тіліне қатысты сөйлеу корпусын және оның мүмкіндіктерімен де танысуға болады. Мысалы, одан сөйлеу жағдайларын табуға мүмкіндік бар. Аталған орыс тілінің сөйлеу корпусын жүзеге асыру әрекетіне Санкт-Петербургтің «Одитек» компаниясы да өз үлесін қосты деуге болады.

Орыс тілінің ISABASE сөйлеу корпусы жайында қысқаша ақпарат.

90-шы жылдардың соңында Ресейдің ғылым академиясының жүйелік талдау институтында тек қана ғылыми мақсаттарға ғана емес, сонымен қатар, Мәскеу мемлекеттік университетінің филология факультетінің сөйлеу корпусымен айналысатын ғылыми топтарының қатысуымен, орыс тіліне қатысты мәтіндерге сөйлеу фрагменттерінің дыбыстық бірліктеріне шартты түрдегі белгіленімдер енгізу арқылы алғашқы орыс тілінің сөйлеу корпусы құрастырылып, қолданысқа ұсынылды. Бұл корпус тек ғылыми зерттеу мақсатына ғана арналған емес, соны-мен бірге дискретті сөйлеу тілінің бірліктерін автоматты түрде тану жүйесін құру мәселесімен де айналысуға арналған еді [4].

RuSpeechatmты сөйлеу корпусын құрастыру жобасы 2000-2001 жж. ИСА РАН тапсырысы бойынша ***Intel*** корпорациясының ғалымдарының күшімен жүзеге асқан болатын. Қазіргі кезде ***RuSpeech*** атты орыс тілінің сөйлеу корпусы ең өкілді деп саналады және басқа тілдердің сөйлеу корпустарын құрастыруға үлгі ретінде пайдалануға бағыт-бағдар беретін аса ыңғайлы корпус. ***RuSpeech*** жобаның ең маңызды нәтижесі деп, сөйлеу корпусын құрудың сыннан өткен технологиясы және осы технологияны қамтамасыз ететін бағдарламалық (компьютерлік) құрал-дар жиынтығы. Бұл жоба орыс тілін автоматты тану жүйесін құруға қатысты ғылыми-зерттеу жұмыстарын және т.б. қажетті деген әректтерді жүзеге асыру шарасында да пайдалануға болады [5]. Сонымен бірге, корпусқа қатысты келесі компьютерлік бағдарламаларды да атап кетуге болады:

– орыс тілінің транскрипторларын автоматтандыруға қатысты бағдарламаны дұрыстау (отладка программы);

– қажетті фонетикалық және статистикалық сипаттамаларына қатысты мәтіндік мәліметтерін жинақтайтын бағдарламаларды құру;

– эксперт-фонетистердің автоматтанған жұмыс орнын құру бағдарламасы;

– диктор сөздерін пакеттік жазу бағдарламасы;

– зерттеу жұмысының негізгі кезеңдерін верификациялайтын (анықтайтын, тексеретін) бірнеше бағдарламаны құру [6].

Соңғы онжылдықта сөйлеу тілін тануға қатысты байқайтынымыз, ол – «қол» ережелері мен алгоритмдеу әдістерінен корпустық модельдеу мен сөйлеу тілін автоматты синтездеу салаларына қарай ауыса бастағандығы. Бұл, әсіресе, сөйлеу тілінің просодикалық сипаттамасын, оның эмоционалды мазмұнын модельдеу үшін аса маңызды, сонымен бірге сөйлеуші дауысының жекелік ерекшелігіндегі еліктеушілігін модельдеу де аса құнды. Сөйлеу корпустары дербес тұрып-ақ ғылыми қызығушылық тудырады, ал әртүрлі тілдердегі дыбыстама сөйлеуді талдау (анализдеу) мен сипаттауға қатысты қажеттілік көптеген ғылыми мәселелерде туындайтыны мәлім.

Сөз соңында, әлемдік басқа тілдермен қатар қазақ тілінің сөйлеу корпустарын ұтымды құрастыру мақсатында төменде сөз болатын жайттарды ескеруіміз қажет деп білем.

Әлем бойынша алғанда, сөйлеу технологияларының екпінді дамуы компьютерлік бағдарлама мен іргелес ғылыми салалардан хабары бар кең түрдегі филолог мамандарын және мақсаттық бағыттағы фонетика мамандарын дайындауға қатаң талап қою қажеттігі. Осыған байланысты Қазақстан Республикасындағы жоғары оқу орындарындағы қазақ тілінің фонетика саласы кафедраларында «Қолданбалы лингвистика (сөйлеу тілінің технологиялары)» атты қосымша ма-мандар дайындау қажет.

Қорыта айтқанда, компьютерлік лингвистика мамандарының тұжырымдауынша, компьютерлік тілдік қор дегеніміз – ғылым адамының өз зерттеу нысанына жаңаша тұрғыда көз салу мүмкіндігі. Мұндай тілдік қор неғұрлым қомақты болса, солғұрлым тіл құрылысының сыры тереңірек ашылады, сөйтіп, зерттелетін нысан жөніндегі түсініктердің аумағы кеңиді, адамның білім өрісіндегі «ақтаңдақтардың» бедер-бейнесі айқындала түседі. Сол сияқты, зерттеуші адамның қалып-қабілеті әлденеше есе артады, шығармашылық қуат көздері ашыла түседі, сөйтіп, бұл жаңа мүмкіндіктер қазақ тілінің жүйелілік қасиеттерін жетілдіруге және тіл жүйесін мұқият тануға жұмсалатыны сөзсіз.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6, №10.

[2] *Захаров В.П.* Корпусная лингвистика [Электронный документ] // http://download.yandex.ru/class/zakharov/CL.

[3] *Баранов А.Н.* Компьютерная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 13-38.

[4] *Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В.* База речевых фрагментов русского языка ISABASE // Сб. «Интеллектуальные технологии ввода и обработки информации». М., Эдиториал УРСС, 1998.

[5] *Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С.* Технология формирования речевых баз данных // Сб. «Организационное управ-ление и искусственный интеллект». М., Эдиториал УРСС, 2003.

[6] *ArlazarovV.L., BogdanovD.S. Krivnova O. F., Podrabinovitch A. Ya.* . Creation of Russian Speech Databases: Design, Processing, Development Tools // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650-656.