А.Қ.ЖҰБАНОВ
А.Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері, филология ғылымдарының докторы, профессор
Алматы қаласы, Қазақстан

ҚАЗАҚ ТІЛІНІҢ СӨЙЛЕУ КОРПУСЫН ҚҰРУДЫҢ ӨЗЕКТІЛІГІ

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.

Тақырыбы: «Цифрлық қазақстан үшін қазақ тілінің сөйлеу корпусын құру өзекті мәселе».

Түйін: 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.

Тема: «Создание корпуса устной речи казахского языка является актуальной проблемой»**.**

Резюме: В статье на мировом уровне дается краткий обзор проблем создания и важности использования корпуса устной речи казахского языка. Кроме того, в статье говориться о важности использования корпуса устной речи казахсого языка как для целей научных исследований, так и для целей создания системы автоматического распознавания единиц дискретной устной речи казахского языка.

Ключевые слова: Корпусная лингвистика, корпус устной речи, компьютерная программа, речевая база, корпусная база, корпусное моделирование, речевой сигнал, языковой ресурс, акустическая вариативность, искуственный интеллект, автоматическое распознавание,

фрейм, синтез речи, дискретная речь, непреривная речь, фонемная/фонетическая/просодическая транскрипция, автоматическое синтезирование, эмоциональное содержание, технология речи.

Theme: “ For Digital Kazakhstan, the creation of a corpus of oral speech of the Kazakh language is a pressing issue”.

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

Қазақ тілін зерттеуде корпустық лингвистика саласының әлемдік дәрежедегі теориялық және практикалық жақтарын зерттеу қажеттігі туындайды. Әлем бойынша мәтін корпустарын құрастыру мен олардың қызметіне қатысты жалпы және нақты мәслелелерге 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 қарастырып 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 зерттеулердің 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 автоматты 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. Заманауи 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 құрылымданған 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 бірнеше 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.

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].

С***өйлеу 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 бірліктерінің статистикалық 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, сонымен бірге дискретті сөйлеу тілінің бірліктерін автоматты түрде тану жүйесін құру мәселесімен де айналысуға арналған еді [4].

RuSpeech атты 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) құралдар 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.

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 мамандар 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.

Әдебиет

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 управление и искусственный интеллект». М., Эдиториал УРСС, 2003.

6. *Arlazarov V.L., Bogdanov D.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.