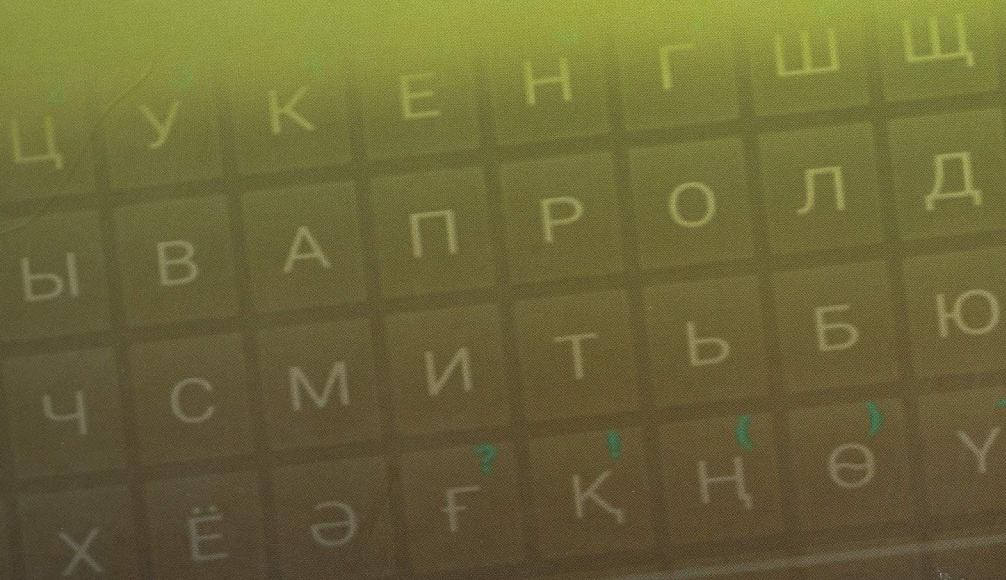


А.Қ. ЖҰБАНОВ, А.Ә. ЖАҢАБЕКОВА

# КОРПУСТЫҚ ЛИНГВИСТИКА



ӨОЖ 811.512.122

КБЖ 81.2 Қаз

Ж 80

А.Байтұрсынұлы атындағы Тіл білімі институтының  
Ғылыми кеңесінде мақұлданған

**Жауапты редактор:** филология ғылымдарының докторы,  
профессор **М. Малбақов**

**Пікір жазғандар:** филология ғылымдарының докторы,  
профессор **Н. Уәли**

филология ғылымдарының  
кандидаты **А. Фазылжанова**

**Жұбанов А.К., Жаңабекова А.Ә.**

**Ж 80 Корпустық лингвистика/ А.К. Жұбанов, А.Ә. Жаңабекова. — Алма-  
ты: «Қазақ тілі» баспасы, 2017. — 336 б.**

ISBN 978-601-7293-43-7

Оқырман назарына ұсынылып отырған оқу құралы корпустық лингвистиканың негізгі ұғымдарымен таныстыра отырып, корпустық технологиялардың негіздерін менгеруге, корпустармен жұмыс істеу дағдысын қалыптастыруға және корпустық лингвистика пәнінің ақпараттық технологиялар қатарынан алатын орнын айқындауға мүмкіндік тудырады.

Мұндай зерттеу жұмысы қазақ тіл білімінде бұрын-соңды қолға алынбағандықтан, біріншіден, зерттеу нәтижесі корпустық лингвистика бойынша алғашқы теориялық тәжірибелердің негізін қаласа, екіншіден, қазақ тілінің ұлттық корпусын құрастыруға бағыт-бағдар береді.

Оқу құралы филология мамандықтарындағы студенттерге, магистранттар мен PhD докторанттарға, компьютерлік лингвистика, корпустық лингвистика пәнінен дәріс оқитын оқытушыларға және ақпараттық технология саласындағы мамандарға арналған, сонымен қатар корпусты құрастырушылар мен оны пайдаланушыларға әдістемелік құрал бола алады.

ӨОЖ 811.512.122

КБЖ 81.2 Қаз

ISBN 978-601-7293-43-7

© Жұбанов А.К., Жаңабекова А.Ә., 2017

© «Қазақ тілі» баспасы, 2017

АЛҒЫ СӨЗ

Оқу құралының мақсаты — оқырманды корпустық лингвистиканың негізгі ұғымдарымен, корпустармен жұмыс істеу жолдарымен, әдіс-тәсілдерімен таныстыру, корпустық технологиялар негіздерін менгеруге мүмкіндік жасау және бұл пәннің басқа да ақпараттық технологиялар қатарынан алатын орнын айқындап беру.

Мақсатқа сай, оқу құралында мынадай міндеттер көзделеді: оқырмандарды корпустық зерттеулер тарихымен таныстыру; оқырмандарды корпустық лингвистиканың тілдік және программалық құралдарын зерттеуге бағыттау; оқырмандарды корпустық лингвистиканың программалық құралдарымен және ақпараттық ресурстарымен жұмыс істеу дағдысын қалыптастыру; оқырмандарды корпустық деректерге негізделген зерттеулермен таныстыру және нақты тілдік материалдармен корпустар арқылы жұмыс істеудің тез және тиімді жүзеге асатындығын көрсету.

Оқу құралы төрт бөлімнен тұрады. «Корпустық лингвистикаға кіріспе» атты *бірінші бөлімде* корпустық лингвистиканың негізгі ұғымдары, терминдері, тіл білімінің жеке саласы ретінде қалыптасу тарихы, тіл білімінің басқа салаларымен байланысы және қолданыстағы корпустар түрлері туралы мәлімет берілген.

«Корпустар құрастыру мәселелері» деп аталатын *екінші бөлімде* лингвистикалық белгіленім (разметка) түрлері, сонымен қатар корпустарды жобалауға, іріктеуге және тілдік материалдарды өңдеуге, белгіленім тәсілдеріне, стандарттауға қатысты технологиялық үдерістер сипатталады.

«Корпустарды пайдалану» атты *үшінші бөлімде* корпустардан қажетті мәліметтерді іздестіруді қамтамасыз ететін корпустық менеджерлер мен корпус материалдары негізінде жүргізілген әртүрлі лингвистикалық зерттеулердің (лексикография, грамматика, курс) нәтижелері сипатталады.

«Қазақ тілінің Ұлттық корпусын құрастыру тәжірибесінен» деп аталатын *төртінші бөлімде* Қазақ тілінің ұлттық корпусын құрастырудағы алғашқы тәжірибе нәтижесі, морфологиялық белгіленімнің лингвистикалық әзірлемесі, мәтіндерге морфология-

лық белгіленім қоюды автоматтандыру, аннотацияланған корпус-стардың ғылыми-практикалық маңызы туралы баяндалады.

Бұл оқу құралын жазуға авторлардың «Компьютерлік лингвистика» мен «Корпустық лингвистика» пәндері бойынша оқыған дәрістері және «Қазақ тілінің Ұлттық корпусын құрастыруға» қатысты теориялық зерттеулерінің (мақалаларының) материалдары мен корпус құрастыру тәжірибесі негіз болды. Сонымен бірге, аталған пән бойынша орыс тілінде жарық көрген филолог-ғалымдар, доцент Виктор Павлович Захаров (Корпусная лингвистика: Учебно-метод. пособие. — СПб., 2005) пен профессор Светлана Юрьевна Богдановамен бірігіп жазған (Корпусная лингвистика: учебник для студентов гуманитарных вузов. — Иркутск: ИГЛУ, 2011. — 161 с.) еңбектері және т.б. осы сала бойынша жазылған ғылыми еңбектер материалдары да негіз болды. Әсіресе, Қазақ тілінің ұлттық корпусы жасалу үстінде болғандықтан, үшінші тарауда ағылшын тілі негізінде жасаған В.Захаровтың корпусты пайдалануға қатысты материалдары түпнұсқа күйінде алынды.

Сонымен бірге осы оқу құралының авторының бірі А.Қ.Жұбановтың қазақ тілін зерттеуге арналған жаңа сала «Қолданбалы лингвистика» бойынша жүргізген ғылыми зерттеу жұмыстары да ұсынып отырған кітапқа теориялық және практикалық жағынан негіз болды деуге болады.

Атап айтқанда, автордың «Қолданбалы тіл білімінің мәселелері» атты монографиялық жұмысы ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институты, ҚР Мәдениет және Ақпарат министрлігі Тіл комитетінің қолдауымен және «Арыс» қорының «Қазақ ғылымының озық үлгілері» атты мегажобасының аясында жарық көрді (Алматы: «Арыс» баспасы, 2008. — 640 бет). Аталған монографияның «Қолданбалы лингвистика: қазақ тілінің статистикасы» деп аталын бірінші бөлімінде қазақ тілі мәтіндеріндегі сөзтұлғалардың лексика-морфологиялық құрылымын статистикалық әдіспен зерттеу үшін шартты белгі-кодты сәйкестендіру (белгіленім) бағдарламасы көрініс тапқан. Мәтін бірліктеріне тән белгі-кодтарды сәйкестендіру «ұялы принципке» негізделіп, негізгі сөз таптары ретінде қазақ тілінің зат есім, етістік, сын есім сөздері тиісті белгі-кодқа сәйкестік кестелер арқылы берілген.

Сол сияқты, аталған монографияның ең өзекті саналатын екінші бөлімі «Основные принципы формализации содержания казахского текста» деп аталып, мұнда «Қазақ тілінің атаушы сөз таптарын семантикалық топтастыру» мәселесі және қазақ мәтінінің семантикалық-мәнді бірліктері мен қисындасу ережелерін тандап алу мәселесі де сөз болады.

А.Қ.Жұбановтың ғылыми зерттеулері қазақ тілінің мәтіндік корпусын жасау мәселесіне теориялық тұрғыдан да, практикалық тұрғыдан да құнды негіз болмақ. Оқу құралын жазу барысында пайдаланылған автордың мұндай ғылыми жұмыстары қазақ тіл білімінде бұрын-соңды қаралмағандықтан, біріншіден, зерттеу нәтижесі корпустық лингвистика бойынша алғашқы теориялық тәжірибелердің негізін қаласа, екіншіден, тілдік материалды модельдеу нәтижесінде көптеген тілдік ақпараттарды алып, оларды тіл білімінің кез келген саласында нақты тілдік дерек ретінде пайдалануға, ұстануға мүмкіндік жасайды.

Қазақ тілінің мәтіндік корпусын зерттеуге арналған бұл жұмыстың нәтижелерін жоғарғы оқу орындарында оқытылатын «Қазақ тілінің функционалды грамматикасы», «Компьютерлік лингвистика», «Корпустық лингвистика», «Морфология», «Мәтін лингвистикасы» т.б. салалар бойынша дәрістер мен семинар сабақтарында пайдалануға болады.

*Филология ғылымдарының докторы,  
профессор  
Мырзаберген МАЛБАҚОВ*

## МАЗМҰНЫ

Алғы сөз .....	3
Кіріспе .....	5

### 1. КОРПУСТЫҚ ЛИНГВИСТИКАҒА КІРІСПЕ

1.1. Корпустық лингвистиканың негізгі ұғымдары, зерттеу нысаны .....	12
1.2. Корпустық лингвистиканы өмірге әкелуге себепші болған бағыт: «картотекадан корпуста».....	24
1.3. Лингвистикалық корпустардың тарихы .....	28
1.4. Корпуста қамтылатын мәтіндер және корпустардың репрезентативтігі .....	51
1.5. Корпустық лингвистика мен дәстүрлі тіл білімінің зерттеу нысандарының айырмашылықтары .....	57
1.6. Корпустық лингвистиканың тіл білімі салаларымен байланысы.....	58
1.7. Корпус мәтіндерінің электрондық жинақтағы мәтіндерден айырмашылығы .....	61
1.8. Корпустарды өртүрлі негізде классификациялау .....	63
1.9. Корпустардың ерекше типтері .....	74
1.10. Қазақ тілінің ұлттық корпусын құрастыру мәселесі .....	84

### 2. КОРПУСТАР ҚҰРАСТЫРУ МӘСЕЛЕЛЕРІ

2.1. Корпус құрастыруды жобалау және технологиялық үдеріс.....	87
2.2. Дерекнаманы іріктеу.....	90

2.3. Табиғи тілді өндеудің негізгі рәсімдері: токендеу (токенизация), леммаға ажырату (лемматизация), стемминг, парсинг .....	92
2.4. Белгіленім ұғымы. Белгіленім түрлері.....	94
2.4.1. Морфологиялық белгіленім.....	101
2.4.2. Корпустағы морфологиялық белгіленімдердің тілдік талдаулармен сабақтастығы .....	105
2.4.3. Зат есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы .....	109
2.4.4. Етістіктің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы.....	114
2.4.5. Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы.....	121
2.4.6. Синтаксистік белгіленім .....	129
2.4.7. Синтаксистік белгіленімнің синтаксистік қатынас теориясымен сабақтастығы.....	138
2.5. Семантикалық белгіленім .....	142
2.5.1. Қазақ тілі лексикасын семантикалық классификациялау ұстанымдары.....	144
2.5.2. Зат есімнің семантикалық классификациясы.....	145
2.5.3. Қазақ тіліндегі етістіктердің семантикалық классификациясы .....	151
2.5.4. Қазақ тілі мәтіндерін туындату моделінде қолданылатын сын есімдердің семантикалық топтары .....	161
2.5.5. Қазақ тілі мәтіндерін туындату моделінде қолданылатын үстеулердің семантикалық топтары .....	170
2.5.6. Қазақ тілі мәтіндерін туындату моделінде қолданылатын есімдіктердің семантикалық топтары.....	172
2.6. Анафоралық белгіленім және просодикалық белгіленім.....	178
2.7. Экстралингвистикалық белгіленім.....	179

2.8. Метабелгіленім енгізілген мәтіндерді арнайы компьютерлік бағдарлама бойынша өңдеу .....	191
2.9. Корпустық лингвистикадағы стандарттау .....	195

### 3. КОРПУСТАРДЫ ПАЙДАЛАНУ

3.1. Корпустық менеджерлер.....	198
3.2. Сауалдар тілі .....	200
3.3. Шығарылым интерфейстері.....	207
3.4. Лингвистикалық емес корпустардың корпустық менеджерлері (WWW) .....	209
3.5. Корпустық зерттеулер.....	213
3.5.1. Корпустарды пайдаланушылар.....	213
3.5.2. Корпустарды пайдалану тәсілдері.....	214
3.5.3. Корпустарға негізделген лексикографиялық зерттеулер .....	216
3.5.4. Корпустарға негізделген грамматикалық зерттеулер .....	234
3.5.5. Корпустарға негізделген дискурстық зерттеу .....	241

### 4. ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫН ҚҰРАСТЫРУ ТӘЖІРИБЕСІНЕН

4.1. Сөздердің морфологиялық құрамын автоматты бөлшектеуге қажетті реестр сөздерді өңдеу .....	257
4.2. Сөзформалар (түрленім) сөздігі – морфологиялық белгіленім қоюдың құралы.....	262
4.3. Тілдің морфологиялық жүйесін автоматты түрде бөлшектеуге қажетті негіз сөздер сөздігін жасаудың лингвистикалық проблемалары.....	266
4.4. Тілдің морфологиялық жүйесін модельдеудегі функционалды қосымшалардың проблемалық мәселелері.....	272

4.5. Қазақ тілінің ұлттық корпусындағы лингвистикалық белгіленім қоюды автоматтандыру.....	281
4.6. Қазақ тілінің мәтіндер корпусындағы омонимдер мәселесі.....	288
4.7. Аннотацияланған корпустардың ғылыми-практикалық маңызы .....	297

ҚОРЫТЫНДЫ .....	305
ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ.....	310
ҚОСЫМША .....	318