Қазақ тілінің Ұлттық корпусын қалыптастыру және жүргізу қағидаларын бекіту туралы

Қазақстан Республикасы Ғылым және жоғары білім министрінің 2025 жылғы 30 сәуірдегі № 227 бұйрығы

"Мәдениет туралы" Қазақстан Республикасы Заңының 24-4 бабының 5-тармағына және "Қазақстан Республикасы Ғылым және жоғары білім министрлігінің кейбір мәселелері туралы" Қазақстан Республикасы Үкіметінің 2022 жылғы 19 тамыздағы № 580 қаулысымен бекітілген Қазақстан Республикасы Ғылым және жоғары білім министрлігі туралы ереженің 15-тармағының 161-1) тармақшасына сәйкес БҰЙЫРАМЫН:

1. Қоса беріліп отырған Қазақ тілінің Ұлттық корпусын қалыптастыру және жүргізу қағидалары бекітілсін.

2. Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Тіл саясаты комитеті Қазақстан Республикасының заңнамасында белгіленген тәртіппен:

1) осы бұйрықтың Қазақстан Республикасы Әділет министрлігінде мемлекеттік тіркелуін;

2) осы бұйрықты Қазақстан Республикасы Ғылым және жоғары білім министрлігінің интернет-ресурсында орналастыруды қамтамасыз етсін.

3. Осы бұйрықтың орындалуын бақылау жетекшілік ететін Қазақстан Республикасының Ғылым және жоғары білім вице-министріне жүктелсін.

4. Осы бұйрық алғашқы ресми жарияланған күнінен кейін қолданысқа енгізіледі.

Министр *С. Нұрбек*

Қазақстан Республикасының
Ғылым және жоғары
білім министрінің
2025 жылғы 30 сәуірдегі
№ 227 бұйрығымен
бекітілген

Қазақ тілінің Ұлттық корпусын қалыптастыру және жүргізу қағидалары

1-тарау. Жалпы ережелер

1. Қазақ тілінің ұлттық корпусын қалыптастыру және жүргізу қағидалары (бұдан әрі – Қағидалар) "Мәдениет туралы" Қазақстан Республикасы Заңының 24-4 бабының 5-тармағына және "Қазақстан Республикасы Ғылым және жоғары білім министрлігінің кейбір мәселелері туралы" Қазақстан Республикасы Үкіметінің 2022 жылғы 19 тамыздағы № 580 қаулысымен бекітілген Қазақстан Республикасы Ғылым және жоғары білім министрлігі туралы ереженің 15-тармағының 161-1) тармақшасына сәйкес

әзірленді және қазақ тілі мен лингвистика саласындағы ғылыми, білім беру және практикалық зерттеулер жүргізуге негіз болатын Қазақ тілінің ұлттық корпусын (бұдан әрі – Ұлттық корпус) құру, толықтыру, өңдеу, сақтау және пайдалануға қойылатын жалпы талаптар мен рәсімдерді белгілейді.

2. Ұлттық корпустың әкімшісі – Қазақстан Республикасының Ғылым және жоғары білім министрлігі.

3. Ұлттық корпусты үйлестіру жұмысын Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетінің "Ахмет Байтұрсынұлы атындағы Тіл білімі институты" шаруашылық жүргізу құқығындағы республикалық мемлекеттік кәсіпорны (бұдан әрі – Институт) жүзеге асырады.

4. Осы Қағидаларда мынадай ұғымдар пайдаланылады:

1) аннотация – мәтіннің авторы, оның тақырыбы, жанры, контексті, құрылымы мен мазмұны, сондай-ақ фонетикалық, морфологиялық, просодикалық, лексикалық, семантикалық, синтаксистік және басқа да лингвистикалық сипаттамалары тәрізді берілген іздеу элементінің дереккөзі туралы ақпаратты қамтитын корпустағы деректердің сипаттамасы;

2) белгіленім – корпус мәтіндеріндегі лингвистикалық және құрылымдық ақпаратты талдау мен өңдеуді қамтамасыз ететін кодтау және жүйелеу әдісі;

3) ғылыми сапа стандарттары – лингвистикалық зерттеулер жүргізу үшін дәлділік, нақтылық және жарамдылықты қамтамасыз ету үшін мәтіндер мен олардың корпустағы аннотациялары сәйкес келетін өлшемшарттар жиынтығы;

4) деректердің бірыңғай форматы – кодтауға, құрылымға және белгілеуге талап қою арқылы өзара дерек алмасу мен талдауды жеңілдететін, қателік қаупін азайта және түзетулер енгізу қажеттілігін төмендете отырып, деректерді үйлесімді және өңдеу мен біріктіруге ыңғайлы ететін стандартты құрылымдар мен ережелер;

5) жанр – материалды талдау мен зерттеу үшін оны жіктеуге көмектесетін, түрін, мақсатын және белгілі бір стилистикалық белгілері бар қалыптасқан формасын айқындайтын мәтіннің категориясы;

6) жиілік – белгілі бір мәтінде немесе корпуста тілдік элементтердің қаншалықты жиі кездесетінін көрсететін сандық көрсеткіш;

7) интерфейс – пайдаланушыға нәтижелерді іздеуге, алуға, қарауға, деректерді сүзуге және талдауға мүмкіндік беретін корпустың бағдарламалық-аппараттық жүйесі;

8) кодтау және жіктеу жүйесі – корпуста деректерді іздеу, сақтау және талдаудың ыңғайлылығын қамтамасыз ете отырып, мәтіндерге бірегей кодтар тағайындауға және оларды негізгі белгілері бойынша жіктеуге мүмкіндік беретін ретті құрылым;

9) корпус – жекелеген ішкорпустардан тұратын және мақсатты лингвистикалық немесе статистикалық талдау жүргізу үшін белгілі бір критерийлер бойынша құрылған Ұлттық корпустың бөлігі;

10) корпустың дерекқоры – корпусқа қосу үшін алдын ала өңделген және белгіленген табиғи тілдегі жазбаша және ауызша мәтіндердің электрондық нұсқалары;

11) қазақ тілінің ұлттық корпусы – қазақ тілінің барлық стиліндегі және жанрындағы мәтіндерді қамтитын, іздеу жүйесімен және мәтінмен жұмыс істеудің басқа да құралдарымен жарақтандырылған ақпараттық-анықтамалық жүйе;

12) мәтіндер – корпустың негізін құрайтын тілдік бірліктер;

13) метабелгіленім – корпусқа енгізілген мәтін мазмұнының тақырыбы, авторы, жанры, мақсаты, стилі және басқа сипаттамалары туралы түпкілікті ақпарат беретін қысқаша дерексөздік сипаттама;

14) морфологиялық белгіленім – корпустағы мәтіндерді автоматты немесе жазбаша түрде морфологиялық сипаттамалармен қамтамасыз ету және олардың грамматикалық белгілерін анықтау процесі;

15) синтаксистік белгіленім – тілдің құрылымы мен оның ережелерін зерттеу үшін сөздер мен олардың грамматикалық рөлдері арасындағы байланыстарды анықтауды қоса алғанда, мәтіндегі сөйлемдердің грамматикалық құрылымын жабдықтау және сипаттау процесі;

16) семантикалық белгіленім – табиғи-тілдік өңдеу, корпустық лингвистика, машиналық оқыту және цифрлық технологиялар әзірлеу мақсатында мәтіндегі сөздердің, сөз орамдары мен сөйлемдердің мағынасын, контекстік реңктерін, сондай-ақ олардың өзара байланыстарын аннотациялау процесі;

17) сөзқолданыс – корпусқа кіретін мәтіндерде сөзді қолдану сипаттамаларының жиынтығы;

18) стиль – мәтіннің мақсаты мен ерекшеліктерін көрсететін реңк, сөздік қор, грамматикалық құрылым және басқа да сипаттамалармен айқындалатын тілдегі жеткізу тәсілі.

2-тарау. Ұлттық корпусты әзірлеу

5. Ұлттық корпусты әзірлеуге, жүргізуге және пайдалануға байланысты барлық жұмыстар деректердің сапасы, қауіпсіздігі мен қолжетімділігінің жоғары деңгейі қамтамасыз етіле отырып, бірыңғай ұлттық тәсіл шеңберінде жүргізіледі.

6. Ұлттық корпус осы Қағидалардың талаптарына сәйкес келетін корпустар мен ішкорпустар есебінен қалыптасады.

7. Ұлттық корпусты қалыптастыру оның қызметінің барлық аспектілерін қамтитын мәтіндерді, тілдің әртүрлі салаларда және уақыт кезеңдерінде қолданылу ерекшеліктерін көрсететін жазбаша дерексөздерден бастап ауызша материалдарға дейін мұқият жинаудан және іріктеуден басталады.

8. Ұлттық корпустағы барлық материалдар ғылыми сапа стандарттарына сәйкес келеді, метадеректері мен дерексөздері болады, грамматикалық және емле қателерінің

бар-жоғына тексеріледі, сондай-ақ бірыңғай кодтау және жіктеу жүйесіне сәйкес өңделеді және реттеледі.

9. Ұлттық корпуста әртүрлі жанрларды, мәтін түрлерін, диалектілер мен стильдерді қоса алғанда лингвистикалық тәжірибелердің алуан түрлілігі көрініс табады, сондай-ақ аймақтық және тарихи ерекшеліктер ескеріледі.

10. Барлық жиналған мәтіндер жанр, стиль, кезең, дереккөз, лексикалық және грамматикалық деңгейлер бойынша жүйеленеді және аннотациясы беріледі.

11. Энциклопедиялық және дереккөздік талдау ескеріле отырып, мәтіндер жиналғаннан және аннотацияланғаннан кейін, ғылыми зерттеулерге қажетті мақсатты эмпирикалық тілдік базалар құруға және тілдік технологиялар мен лексикографиялық жобаларды әзірлеуде қолданылатын тіл модельдерін әзірлеуге мүмкіндік беретін лингвистикалық өңдеу және деректерді талдау әдістері қолданылады.

12. Ұлттық корпустың функциялары:

1) Метабелгіленім мәтінді және оның сипаттамаларын, соның ішінде авторларды, тақырыпты, жанрды, дереккөзді, әзірленген күнін, мақсатты аудиторияны, стильді, мәтіннің өлшемін, форматын, құрылымдық ерекшеліктерін егжей-тегжейлі сипаттайтын ақпаратты береді;

2) Ұлттық корпустың цифрлық құрылымы мәтіндерді цифрлық форматта жүйелеуді қамтамасыз етеді, бұл заманауи автоматтандырылған талдау құралдарын пайдалануға мүмкіндік береді;

3) Ұлттық корпустың көпфункционалдығы оны лингвистикалық зерттеулер, лексикография, тілдерді оқыту және компьютерлік лингвистика және нейролингвистикалық бағдарламалау сияқты әртүрлі салаларда қолдануға мүмкіндік береді;

4) Ұлттық корпустың қолданбалы міндеттері мәтінді өңдеуді автоматтандыруға, тілдік модельдердің сапасын жақсартуға, мәтінді талдаудың тиімді құралдарын әзірлеуге, тілдерді оқытуға арналған қосымшалар жасауға, машиналық аудармаға және жасанды интелектті әзірлеуге мүмкіндік береді.

13. Ұлттық корпустың құрылымы:

1) Ұлттық корпус әрқайсысы нақты бағыттарды қамтитын жекелеген корпустардан және ішкорпустардан тұрады;

2) Барлық корпустар және ішкорпустар Әкімшінің келісімімен Институт белгілеген Ұлттық корпустың техникалық, функционалдық және операциялық сипаттамаларына сәйкес келеді, бұл пайдалану барысында біркелкілікті, үйлесімділікті және қатесіз интеграцияны қамтамасыз етеді;

3) Ұлттық корпустағы барлық мәтіндер бірегей форматтарда сақталады, сондай-ақ API кіріктірілуі қамтамасыз етіледі;

4) Мәтіндік деректер ғылыми жұмыстарды, көркем шығармаларды, публицистика, іскерлік құжаттама, ауызекі сөйлеу және сандық мәтіндерді қоса алғанда жанрлар мен стильдердің кең ауқымын қамтиды;

5) Лексикалық ақпарат сөздер, олардың формалары, мағыналары, үйлесімділігі және сөз таптары мен тақырыптық бағыттар бойынша жіктелген жиілік, стилистика, синонимдер, антонимдер тәрізді метадеректерді қамтиды;

6) Грамматикалық ақпарат сөйлем құрылымы, сөз таптары, синтаксистік байланыстар және морфологиялық сипаттамалар талданған синтаксис, морфология және сөзжасам сипаттамаларын қамтиды;

7) Стилистикалық ақпарат мәтін ерекшеліктерінің сипаттамасы, стильдер арасындағы айырмашылық, әртүрлі жанрлардағы тілдік талдау және мәтіндерді ресми және бейресми стильдер, жанрлар мен оларға тән сипаттамалар тәрізді категориялар бойынша жіктеуді қамтиды;

8) Ұлттық корпуста машиналық аударма, контрастивтік лингвистика саласында зерттеулер жүргізуге және тілдік байланыстардың әсерін зерделеуге мүмкіндік беретін параллель мәтіндерді қамтитын кросс-тілдік байланыс бар;

9) Әрбір мәтін сәйкестендіру деректерін, библиографиялық деректерді, жанрлық-стилистикалық сипаттамаларды, лингвистикалық деректерді, контекстік деректерді, техникалық деректерді және аннотациялық деректерді қоса алғандағы метадеректермен қамтамасыз етіледі;

10) Мәтіндік деректермен бірге жүретін медиафайлдар мультимодальды ақпарат контекстінде тілдің әртүрлі аспектілерін талдауға қызмет етеді;

11) Интерфейс деректерді іздеу, талдау және аннотациялау үшін пайдаланушының Ұлттық корпуспен ыңғайлы өзара әрекеттесуін қамтамасыз етеді;

12) Іздеу жүйесі пайдаланушыларға әртүрлі критерийлер негізінде Ұлттық корпустың дерекқорынан деректерді іздеуге және алуға мүмкіндік береді;

13) Ұлттық корпустың ресми платформасы нақты уақытта сөзқолданыс санын бақылауға мүмкіндік береді;

14) Деректер құрылымы қажет болған жағдайда Ұлттық корпусқа жаңа санаттар мен компоненттерді қосуға мүмкіндік береді.

3-тарау. Ұлттық корпусты жүргізу

14. Ұлттық корпус тілдегі өзгерістер мен әртүрлі салалардағы өзекті жетістіктерді көрсететін жаңа материалдардың енгізілуі арқылы үздіксіз жаңартылып, кеңейтіледі, архивтік нұсқаларының қолжетімдігі қамтамасыз етіледі.

15. Деректер базасының сапасы мен өзектілігін арттыру үшін халықаралық ғылыми және лингвистикалық ұйымдармен өзара іс-қимыл қамтамасыз етіледі.

16. Ұлттық корпустың тиімді жұмыс істеуін үйлестіру үшін қажетті білімі мен тәжірибесі бар бейінді мамандардан тұратын жұмыс тобы құрылады.

17. Ұлттық корпусқа жүктелген барлық материалдар форматтау, емле, пунктуация, стиль және құрылым талаптарына сәйкес келеді.

18. Ұлттық корпусқа қосу үшін материалдар ұсынатын барлық қатысушылар олардың сапасын, дәлдігін, шынайылығын, өзектілігін қамтамасыз етеді.

19. Деректердің жоғары сапасын қамтамасыз ету үшін жүйелі түрде мониторинг жүргізіледі.

20. Ұлттық корпус лексикографиялық және грамматикалық жобаларды, ғылыми және білім беру материалдарын әзірлеуге, сондай-ақ бағдарламалық өнімдерді жасауға негіз болады.

21. Ұлттық корпусты жүргізу кезінде ақаулар немесе қателер орын алған жағдайда Институт оларды жою үшін шаралар қабылдайды, сондай-ақ Ұлттық корпустың тұрақты және үздіксіз жұмысын қамтамасыз ету үшін бағдарламалық жасақтаманы жаңартып отырады.

22. Ұлттық корпус тілдегі өзгерістердің, жаңа тенденциялардың, сондай-ақ жаңа жанрлардың, стильдер мен бағыттардың пайда болуы есебінен кеңейеді, бұл оның өзектілігі мен тілдік инновацияларға сәйкестігіне кепілдік береді.

23. Егер оларды пайдалану авторлық құқықты бұзбаса және әділ пайдалану нормаларына сәйкес келсе, Ұлттық корпусқа ашық және қолжетімді көздерден алынған материалдар енгізілуі мүмкін.

24. Ұлттық корпус, корпустар пен ішкорпустар жалпыға қолжетімді және барлық мүдделі тараптарға пайдалануға ұсынылады.

25. Ұлттық корпусты жүргізу кезінде кемсітушілік мазмұнға жол бермілмейді және инклюзивтілікті қамтамасыз етуді қоса алғанда, этикалық қағидаттар ескеріледі.