



ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ



Студенттер мен жас ғалымдардың
«ҒЫЛЫМ ЖӘНЕ БІЛІМ - 2014» атты
IX халықаралық ғылыми конференциясы

IX Международная научная конференция
студентов и молодых ученых
«НАУКА И ОБРАЗОВАНИЕ - 2014»

The IX International Scientific Conference for
students and young scholars
«SCIENCE AND EDUCATION-2014»

2014 жыл 11 сәуір
11 апреля 2014 года
April 11, 2014



**ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ**

**Студенттер мен жас ғалымдардың
«Ғылым және білім - 2014»
атты IX Халықаралық ғылыми конференциясының
БАЯНДАМАЛАР ЖИНАҒЫ**

**СБОРНИК МАТЕРИАЛОВ
IX Международной научной конференции
студентов и молодых ученых
«Наука и образование - 2014»**

**PROCEEDINGS
of the IX International Scientific Conference
for students and young scholars
«Science and education - 2014»**

2014 жыл 11 сәуір

Астана

УДК 001(063)
ББК 72
Ғ 96

Ғ 96

«Ғылым және білім – 2014» атты студенттер мен жас ғалымдардың IX Халықаралық ғылыми конференциясы = IX Международная научная конференция студентов и молодых ученых «Наука и образование - 2014» = The IX International Scientific Conference for students and young scholars «Science and education - 2014».
– Астана: <http://www.eni.kz/ru/nauka/nauka-i-obrazovanie/>, 2014. – 5831 стр.
(қазақша, орысша, ағылшынша).

ISBN 978-9965-31-610-4

Жинаққа студенттердің, магистранттардың, докторанттардың және жас ғалымдардың жаратылыстану-техникалық және гуманитарлық ғылымдардың өзекті мәселелері бойынша баяндамалары енгізілген.

The proceedings are the papers of students, undergraduates, doctoral students and young researchers on topical issues of natural and technical sciences and humanities.

В сборник вошли доклады студентов, магистрантов, докторантов и молодых ученых по актуальным вопросам естественно-технических и гуманитарных наук.

УДК 001(063)
ББК 72

ISBN 978-9965-31-610-4

©Л.Н. Гумилев атындағы Еуразия ұлттық
университеті, 2014

ҰЛТТЫҚ МӘТІНДІК КОРПУСТАРҒА ШОЛУ ЖАСАУ**Ергеш Ж.Ж., Ергеш М.Ж.**zharas.yergesh@mail.ru

Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық технологиялар факультеті, «6М060200-Информатика» мамандығының магистранттары

Ғылыми жетекшісі – Бекманова Г.Т

Корпус – белгілі бір тілде электрондық нұсқада мәтіндерді жинауға негізделген ақпараттық-анықтамалық жүйе. Ұлттық мәтіндік корпус тілдің белгілі бір сатысында (немесе сатыларында) оның жүзеге асуы және жанрлардың, үгілердің, аймақтық және әлеуметтік нұсқалардың көптүрлілігі және тағы басқалар.

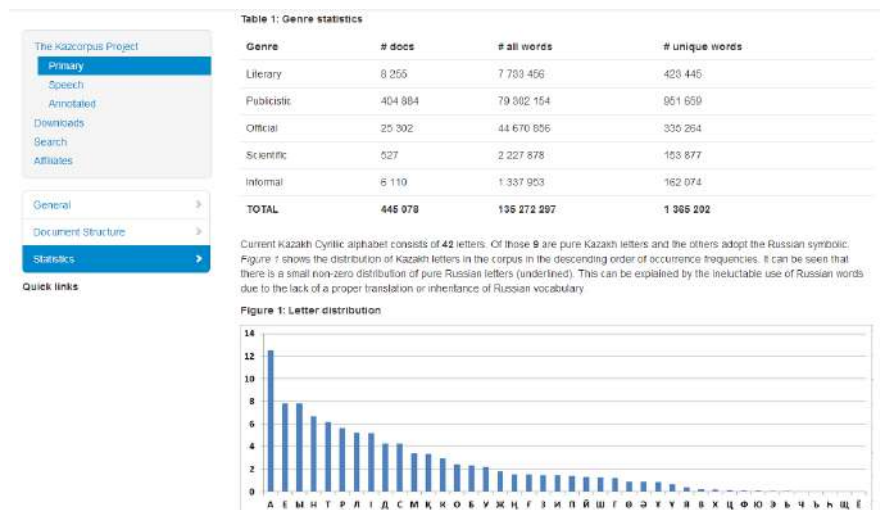
Ұлттық корпус ғылыми зерттеулер үшін және тілді үйрету мақсатымен лингвист (қазіргі таңда тез қарқынмен тіл саласында дамып келе жатқан *корпустық лингвист* деп аталатын маман) мамандардың көмегімен құрылады. Әлемдегі кең таралған тілдердің көпшілігінде өздерінің ұлттық корпустарына ие (мәтіндердің ғылыми өңделуінің деңгейіне және толыққандылығына байланысты). Жалпыға белгілі үлгілердің бірі *Британ ұлттық корпусы (BNC)*: оған қарап, қазіргі кезде көптеген басқа корпустар бағыт-бағдар алуда.

Ұлттық мәтіндік корпусының екі маңызды ерекшелігі бар. Біріншіден, ол өкілділік сипатында болады немесе қалыпқа келтірілген мәтіндер құрамымен. Бұл дегеніміз берілген тілде корпус мүмкіндігінше жазбаша және ауызша мәтіндердің барлық түрлерін қамтиды (әр түрлі жанрдағы көркем әдебиет, публицистикалық, оқу, ғылыми, ауызша, диалектілі, іскерлік және тағы басқа) және бұл мәтіндердің барлығы сәйкес уақыт аралығына байланысты тілдің құрамында мүмкіндігінше пропорционал түрде корпусқа енеді. Екіншіден, корпус оның құрамына кіретін мәтіндердің қасиеті туралы ерекше қосымша ақпаратқа ие болады.

Қазақ тілінде мәтіндік корпустар

Қазақ тілі үшін Kazakh Language Corpus корпусы бар, ол іздеу механизмін, жүктеу үлгілерін, құжатнамалар және кейбір қызықты статистикалық мәліметтер мен ақпараттан тұрады. Бұл жоба ашық түрдегі қазақ тілі корпусының алғашқы бастамасы. Бұл корпус 5 статистикалық жанр бойынша, оның ішінде көркем әдебиет, іскерлік, ғылыми, публицистикалық және формальды емес түрлерінен 135 миллиондан астам сөздерден құралған (2-сурет). Қазақ тілінің бұл корпусында келесі ішкі корпустар қамтылған: сегменттелген құжаттан тұратын, XML -де кодталған, морфологиялық және синтаксистік белгілеулерден тұратын ішкі корпус; ауызша айтылған ішкі корпус; ауызша мәліметтермен аннотацияланған ішкі корпус [1].

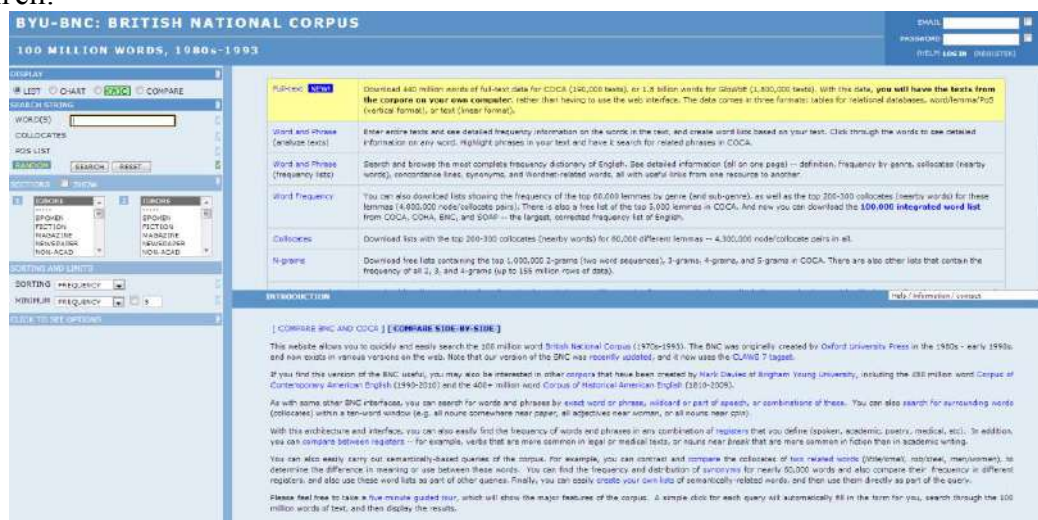
Сонымен қатар [2] жұмыста А.Байтұрсынов атындағы тіл білімі институтында «Қазақ тілінің аннотацияланған Ұлттық корпусы» жобасы орындалып жатыр. Жобаның мақсаты, қазақ әдебиеті тіліндегі әр түрлі функционалды стильді қамтитын ұлттық корпусының терең лингвистикалық мәліметтермен аннотацияланған мәтіндерін құру.



1-сурет. Қазақ тілі корпусының интерфейсі

Басқатілдердің мәтіндік корпустарына шолу

Кең таралған үлгілердің ішінде Британ ұлттық корпусы (BNC): қазіргі таңдағы көптеген корпусстар соған бағытталған [3]. 2-суретте Британ ұлттық корпусының интерфейсі көрсетілген.



2-сурет. Британ ұлттық корпусының интерфейсі

Ағылшын тілінде біршама корпусстар бар. Бұл корпусстар тізімінде: қазіргі заманғы ағылшын тілінің корпус (COCA), тарихи америкалық ағылшын корпусы (COHA), Ағылшын тілінің халықаралық корпусы (ICA), Браун корпусы, TIME журналының корпусы және тағы басқалар [4]. Бұл корпусстардың барлығының интерфейсі бір, онда әр түрлі параметр арқылы іздеу функциясы қамтылған. Төменде жоғарыда аталған корпусстар туралы қысқаша мәлімет келтіреміз.

Браун корпусы 1960 жылдары құрылған тарихтағы алғашқы корпус. Корпус 500 прозалық фрагменттерден құралып 2000 сөзді қамтиды, 1961 жылы АҚШ-та жарияланған мәтіндерден алынған. 1970 жылдың соңында корпус сөз таптары белгілеуімен және сөздердің морфологиялық қасиеттеріне байланысты мәліметтермен толықтырылды. Корпусты көлемі 1 млн сөздердің қолданысынан тұрады.

Қазіргі заманғы ағылшын тілі корпусы (COCA) – ағылшын тілінің кең таралған, үлкен, балансталған корпусы және қазіргі таңда 445 млн сөздердің қолданысын қамтитын ең алып ағылшын тілінің жалғыз корпус. Корпус бірдей пропорциялы ауызша сөздерден, көркем әдебиеттен, журнал мақалаларынан, газеттер және ғылыми мәтіндерден құралған. Әр жыл

сайын 1990 жылдан 2009 жылға дейін 20 миллион көлеміндегі мәтін сөздерімен таралған. Корпус жылына бір немесе екі рет жаңартылып отырады.

Тарихи америкалық ағылшын корпусы (СОНА) – 1810 жылдан 2009 жылға дейінгі аралықтағы мәтіндерді қамтитын 400 миллион сөздерден тұратын тарихи корпус. Корпустың авторы Марка Дэвис.

Ағылшын тілінің халықаралық корпусы (ІСА) ағылшын тілінің әр түрлі нұсқасында қолданылуын көрсетеді (Австралия, Ұлыбритания, Гонконг, Индия, Ирландия, Канада, Кения, Малайзия, Жаңа Зеландия, Сингапур, АҚШ, Танзания, Филиппин, Шри-Ланка, Оңтүстік Африка, Ямайка). Әрбір ішкі корпус әр түрлі типтегі жазбаша және ауызша мәтіндерді қамтиды. Қазіргі таңда толығымен Британдық компонент корпусымен (ICE – GB) дайындалған, оның мәтіндері морфологиялық және синтаксистік белгілеулерден құралған. Әрбір ұлттық ішкі корпустың көлемі – 1 миллион сөздердің қолданысынан тұрады.

TIME журналының корпусы – 100 миллион сөз, 1923 – 2006 жылдар аралығында басып шығарылған журналдардың мәтіндерін қамтиды.

Славяндық тілдердің ішіндегі корпустардан Прага қаласындағы Карл университетінде құрылған Чехиялық ұлттық корпус ерекше орынға ие [5].

Орыс тілінің ұлттық корпусы XVIII ғасырдың ортасы мен ХХІ ғасырдың басындағы уақыт аралығындағы сөздерді қамтиды: бұл уақыт аралығы алдыңғы дәуірлердегі тілдерден құралған, және әр түрлі қазіргі заманғы социоллингвистік нұсқадағы мәтіндерден де қамтиды – әдеби, ауызша, қарапайым ауызша және диалект сөздер. Корпус мәдени мәнге ие болған және тіл жағынан алып қарағанда белгілі бір мәнге болатын көркем әдебиет шығармаларының түпнұсқаларынан тұрады. қазіргі таңда орыс тілінің ұлттық корпусы мына ішкі корпустардан құралған: негізгі корпус, синтаксистік (терең аннотацияланған) корпус, газеттік корпустар (қазіргі заманғы бұқаралық ақпарат құралдары корпусы), параллельдік корпус, ауызша тіл корпусы, диалектілі сөздер корпусы, поэтикалық мәтіндер корпусы, орыс тілін оқып-үйрететін корпус, акцентологиялық корпус (орыс тілі екпіні тарихының корпусы), мультимедиялық корпус [6]

Сонымен қатар, 50 миллион сөзден тұратын түрік тілінің ұлттық мәтіндік корпусы бар. Мұнда 20 жылды қамтитын (1990-2009 ж) әр түрлі жанрадағы мәтіндік мәліметтердің үлгілері жиналған [7,8].

Қорытынды

Ұлттық корпус ең алдымен тілдің лексикасы мен грамматикасына зерттеу жұмыстарын жүргізу үшін арналған, дәлірек айтқанда үлкен емес аралықта – бір немесе екі жүзжылдықтардағы үзіліссіз тілдік өзгерістер процессін қарастырады. Көрсетілген облысқа байланысты (лексика, грамматика, акцентология, тіл тарихы) жан-жақты анықтама беру. Қазіргі заманғы компьютерлік технология үлкен көлемді мәтіндердің лингвистикалық өңделу процесін жеңілдетеді және тездетеді. Қазақ тілді бағдарламаларды жасауда тілдік салада көп көмегін тигізеді .

Қолданылған әдебиеттер тізімі

1. <http://kazcorpus.kz>
2. Жұбанов А.Қ., Жаңабекова А.Ә., Құлманов С. Филологтар қауымдастығы мен корпустық лингвистика орталығын құру – қазақ тілінің ұлттық корпусын жасаудың алғышарты // «Түркі тілдерін компьютерлік өңдеу» атты І халықаралық конференция еңбектері.- Астана, 2013.- Б. 60-67
3. <http://www.natcorp.ox.ac.uk/>
4. <http://corpus.byu.edu/>
5. <http://ucnk.ff.cuni.cz/>
6. <http://www.ruscorpora.ru/corpora-structure.html>
7. <http://www.tnc.org.tr/index.php/en/homepagee>