

К. Дуйсенов

А. Байтұрсынұлы атындағы Тіл білімі институтының
кіші ғылыми қызметкері, педагогика ғылымдарының магистрі

ЛАТЫН ГРАФИКАСЫ НЕГІЗІНДЕГІ ЖАҢА ҚАЗАҚ ӘЛІПБИІ ЖӘНЕ UNICODE ЖҮЙЕСІ

Аннотация. Мақалада латын графикасы негізіндегі қазақ әліпбиін енгізудің кейбір техникалық аспектілері қарастырылады. Әліпбиде базалық латынның 26 таңбасын ғана пайдаланудың өзектілігі талданады. Техникалық тұрғыдан кеңейтілген латын таңбаларын пайдаланудың мүмкіндіктері қарастырылады. Unicode – таңбаларды кодтаудың халықаралық стандартына шолу жасалады. Unicode-тың нұсқалары, басылымдары, қамтитын тілдері мен жазу жүйелері және жалпы таңбалар жиынтығы сипатталады.

Тірек сөздер: латын графикасы, жаңа қазақ әліпбиі, Unicode, ISO/IEC 10646, UTF-8, латын таңбалары.

К.Дуйсенов

Младший научный сотрудник Института языкознания им. А.Байтұрсынова,
магистр педагогических наук.

НОВЫЙ КАЗАХСКИЙ АЛФАВИТ НА ОСНОВЕ ЛАТИНСКОЙ ГРАФИКИ И СИСТЕМА UNICODE

Аннотация. В статье рассматриваются некоторые технические аспекты внедрения казахского алфавита на основе латинской графики. Анализируется актуальность использования в алфавите только 26 символов базовой латыни. Рассматриваются возможности использования технически расширенных латинских символов. Сделан обзор международного стандарта кодирования Unicode – символов. Описываются версии, публикации, языки, системы записи и общий набор символов Unicode.

Ключевые слова: латинская графика, новый казахский алфавит, Unicode, ISO/IEC 10646, UTF-8, латинские символы.

K.Duysenov

A. Junior researcher at the Institute of Linguistics named after A. Baitursynov,
Master of Pedagogical Sciences.

THE NEW KAZAKH ALPHABET BASED ON THE LATIN SCRIPT AND THE UNICODE SYSTEM.

Annotation. The article discusses some technical aspects of the introduction of the Kazakh alphabet based on Latin graphics. The relevance of using only 26 characters of

basic Latin in the alphabet is analyzed. The possibilities of using technically extended Latin characters are considered. Unicode-overview of the international character encoding standard. It describes the Unicode versions, publications, languages, and writing systems, and the common character set.

Keywords: Latin script, new Kazakh alphabet, Unicode, ISO/IEC 10646, UTF-8, Latin characters.

Әліпбиіміздің латын қарпіне көшуіне байланысты әлеуметтік желі мен БАҚ-та пікірталас, ақпарат кеңістігінде толқу туғызып жатқан мәселенің ең бастысы әліпбидегі әріп құрамға қатысты болып отыр. Көптеген әліпби нұсқасы көрсетіліп, түрлі топ өз идеясын ұсынып жатыр.

Осы әліпби жобаларын қарастырғанда оларға тән бір ортақ ерекшелік байқалды. Бұл құбылысқа үйреніп кеткеніміз соншалық тіпті мән бермейміз. Әлеуметтік желі, сайт, түрлі платформаларда ұсынған жобаларда латын әліпбиі таңбалары сол қалпы көрініс береді. Ғаламтор таңбаларды бірден таниды, түрін бұзбай коммуникация құралдарының монитормына шығарады. Қалайша сайттар әлемнің әртүрлі жазуындағы алуан әріп таңбаларын таниды? Әріптер әртүрлі әлеуметтік желілер мен мессенджерлерде бірдей жинақы көрінеді, диакритикалық белгілерге де қарамай еш қиындықсыз форматталады. Диакритикалық белгі түгіл қытай иероглифтерін, армян, грузин, еврей, ежелгі грек, копт сияқты көнеден келе жатқан жазу таңбаларын да бірден ажыратады. Осы сұрақтардың жауабын табу мақсатында ізденісімізді сайттардың кодын ақтарудан бастадық.

Ғаламтор парақтарында символдарды тану үшін “charset” (‘character set’ – таңбалар жиынтығы) атрибуты жауап береді. Көп жағдайда “charset” атрибутының жанында UTF-8 (UTF-16, UTF-32) таңбалар кодтауын пайдаланады. UTF-8 (Unicode Transformation Format) – таңбаларды кодтаудың ресми бекітілген және өзіне 1 112 064 таңбаға дейін сыйдыра алатын формат. UTF-8 негізі ISO/IEC 10646 Әмбебап Кодталған Таңбалар Жиынтығы Халықаралық стандартынан [1] алынады. Стандарт ақпараттық технологиялар саласында қолданылады. ISO/IEC 10646 стандарты өз кезегінде Unicode консорциумының кестелеріне сүйенеді.

Unicode жайында біраз ақпарат бере кеткен жөн болмақ. Unicode – 1991 жылы технологиялық компаниялар арасында таңбалардың кодталуын стандарттау бойынша жұмыс тобы негізінде құрылған коммерциялық емес консорциум. Консорциум 1991 жылдан бастап унификацияланған және кодталған таңбалардың кестесі мен әлем тілдерінің жазу жүйелері таңбаларын жариялап келеді. Сондай-ақ Unicode (көне түркі, мысыр, авеста, руна жазуы, сына жазу секілді) қолданыстан шыққан жазуларды (ескі сирия, манихей, ронго, финикия және т.б.), өлі тілдердің таңбаларын, брайль іздері, валюта таңбаларын, техникалық символдарды (маджонг, ойын картасы, домино және т.б.), ойын таңбаларын, емоджилерді кодтаумен айналысады. Осы уақытқа дейін Unicode кестелерінің 21 басылымы жарияланды (кестені қараңыз).

2017 жылы маусымда жарияланған Unicode 10.0 [2] нұсқасында барлығы 136 755 таңба бар. Оның ішінде 954 [3] латын таңбасы бар. 128 US-ASCII және 826 кеңейтілген латын таңбасы. Сондай-ақ 288 фонетикалық таңба бар. Unicode базасындағы таңбалардың алуан түрлілігі мен кеңдігі туралы мұндай ақпараттан хабардар болған кез келген адам үшін 2017 жылдың 11 қыркүйегінде ҚР Парламентінде қазақ тілінің жаңа латын әліпбиін ұсынған Ш.Шаяхметов атындағы Тілдерді үйлестіру орталығының директоры Е. Тілешовтің “смартфондар мен өзге де жазба құралдары бізге түрлі елден келеді, онда латынның 26 ғана әрпі бар” [4] дегені не әдейі ақпаратты бұрмалау, не білместіктен туындаған қате пікір болып табылады. Себебі қазіргі ақпараттық

технологиялар саласындағы стандарттардың икемділігі мен ақпаратты өңдеу, жанартудың алуан түрлі мүмкіндіктері зор, бұларды жоғарыда келтірілген мәліметтер, тіпті осы мақаланы оқып отырған сайттың коды да дәлелдеп отыр. Мұның бәрі Е.Тілешов мырзаның өз жобасының артықшылығы ретінде келтірген пікірін жоққа шығарады.

Қазақ тілінің латын әліпбиі жобаларымен танысудан туындаған бақылауларымызға жауап іздеу бізді IT-кеңістіктегі латынграфикалы таңбаларға қатысты техникалық құжаттарды, әртүрлі халықаралық стандарттарды жан-жақты зерт-теуге итермеледі. Әр құжатпен танысып отырып, күнделікті смартфон бетінде көретін түзу де әдемі әріптердің артында үлкен техникалық негіз (background) бар екеніне көзіміз жетті. Unicode консорциумы 1991 жылдан бері бүкіл таңбаларды стандарттап дайындап, өте жемісті жұмыс жасағанын көрдік. Демек, Unicode базасында жоқ деп, пернетақтадағы 26 қаріппен шектеліп, төл дыбыс және төл жазуымыздың принциптерінен бас тартып, әліпбиіміздің тілдік тұрпатынан айырылып қалуға еш негізі жоқ. Сондықтан өзі стандартталып, өзі кодталып қойған Unicode базасындағы латын таңбаларынан тіліміздің ішкі талаптарына жауап беретін әліпби жасамасақ, бізге сын. Unicode кестелеріндегі бас әріп және кіші әріптің 954 латын таңбасы ішінен төл дыбыстарымызды қамтитын, тілдің ішкі заңдылығына сәйкес келетін ыңғайлы әрі ұтымды әліпбиге таңба табылар деген үміт зор.

Мақалада келтірілген барлық дерек ашық ақпарат көздерінен алынды.

Unicode нұсқалары

Нұсқа	Уақыты	ISO/IEC 10646 редакцияларындағы көрінісі	Тілдер/Жазу жүйелері	Таңбалар	
				Барлығы	Ескерілетін мәлімет
1.0.0	1991 қазан		24	7 161	Араб, Армян, Бенгали, Бопомофо, Кирил, Деванагари, Грузин, Грек және Копт, Гуджарати, Гурмухи, Хангул, Еврей, Хирагана, Канада, Катакана, Лао, Латын, Малайлам, Ория, Тамил, Телугу, Тай, Тибет таңбалары қосылды.
1.0.1	1992 маусым		25	28 359	20 902 унификацияланған Қытай, Жапон, Корей (бұдан әрі – ҚЖК) идеографы қосылды.
1.1	1993 маусым	ISO/IEC 10646-1:1993	24	34 233	Бастапқы 2 350 Хангул таңбасына 4 306 астам таңба қосылды. Тибет алынып тасталды.
2.0	1996 шілде	ISO/IEC 10646-1:1993 қосымша толықтырулар 5, 6 және 7	25	38 950	Бастапқыдағы Хангул таңбалары өшірілді, ал жаңа 11 172 таңба жаңа орынға көшірілді. Тибет жаңа таңбалар тізбегімен артқы жаңа орынға ауысты.
2.1	1998 мамыр	ISO/IEC 10646-1:1993 қосымша толықтырулар	25	38 952	Еуро таңбасы және объектінің орнын ауыстыру таңбасы қосылды.
3.0	1999 қыркүйек	ISO/IEC 10646-1:2000	38	49 259	Чироки, Эфиоп, Кхмер, Моңғол, Бирма, Огам, Руникалық, Синхала, Сирия, Тана, Біріккен Канада аборигендері таңбалары, И таңбалары мен Брайль іздері қосылды.

3.1	2001 наурыз	ISO/IEC 10646-1:2000 ISO/IEC 10646-2:2001	41	94 205	Десерет, Готикалық және ескі Италия жазуы, Батыс музыкалық және Византия музыкалық таңбалары мен 42 711 ҚЖК таңбасы қосылды.
3.2	2002 наурыз	ISO/IEC 10646-2:2001	45	95 221	Филиппин жазулары: Бухид, Хануно, Тагалог, Тагбануа қосылды.
4.0	2003 сәуір	ISO/IEC 10646:2003	52	96 447	Кипр жазулары, Лимбу, Linear B, Осман (Сомали жазуы), Шави, Тай Ли, Угар жазуы және Гексаграмма таңбалар қосылды.
4.1	2005 наурыз	ISO/IEC 10646:2003 қосымша толықтырулар 1	59	97 720	Бугин, Глаголица, Харошти, Жаңа Тай Лу, Ескі Парсы, Силюти, Нагри, Тифна қосылды, Грек пен Копт бөлінді, Ескі грек сандары мен музыкалық таңбалар қосылды.
5.0	2006 шілде	ISO/IEC 10646:2003 қосымша толықтырулар 1, 2	64	99 089	Балин, Сына жазу, Нко, Фагспа, Фонетикалық таңбалар қосылды.
5.1	2008 сәуір	ISO/IEC 10646:2003 қосымша толықтырулар 1, 2, 3 және 4	75	100 713	Кари, Чам, Кайа Ли, Лепча, Лесия, Лидия, Ол Чики, Реджанг, Саураштра, Судан мен Вай қосылды. Фаистос дискісі, Малжонг тастары, Домино тастары таңбалары қосылды.
5.2	2009 қазан	ISO/IEC 10646:2003 қосымша толықтырулар 1, 2, 3, 4, 5 және 6, теңге таңбасы	90	107 361	Авеста, Бамум, Мысыр, Империялық Арамей, Көне Пехлеви, Көне Парфия, Ява, Кайти, Ли Су, Митей Майек, Ескі Оңтүстік Араб, Көне Түркі, Самаритан, Тай Там, Тай Виет жазулары қосылды.
6.0	2010 қазан	ISO/IEC 10646:2010 қосымша үнді рупий таңбасы	93	109 449	Багақ, Брахма, Мандайқ, ойын қарталары таңбалары, көлік және карта таңбалары, алхимия таңбалары, эмотикон мен емоджи, 222 унификацияланған ҚЖК таңбасы қосылды.
6.1	2012 қаңтар	ISO/IEC 10646:2012	100	110 181	Чакма, Меройтик қиғаш, Миао, Шарала, Сора, Сомпенг, Такри қосылды.
6.2	2012 қыркүйек	ISO/IEC 10646:2012 қосымша түрік лирасының таңбасы	100	110 182	Түрік лирасының таңбасы.
6.3	2013 қыркүйек	ISO/IEC 10646:2012 қосымша алты таңба	100	110 187	5 қосбағытты форматтау таңбасы.
7.0	2014 маусым	ISO/IEC 10646:2014 қосымша толықтырулар 1 және 2, рубль таңбасы	123	113 021	Басса Ва, Кавказ Албан, Дуплоян, Елбасан, Гранта, Ходжжи, Кудоуади, Linear A, Махаджани, Манихей, Менде Кикакуй, Моди, Мро, Набата, Ескі Солтүстік Араб, Ескі Перм, Пахау, Хмонг, Палмир, Пау Син Хау, Орта Пехлеви, Сиддам, Тирхута, Уаранг Сити, Динбатс.

8.0	2015 маусым	ISO/IEC 10646:2014 қосымша толықтырулар 1, лари таңбасы, 9 унификацияланған ҚЖК (Қытай, Жапон, Корей) таңбасы, 41 емоджи таңбасы	129	120 737	Ахом, Анадолы Иероглифтері, Хатран, Мултани, Ескі Мажар, Таңба жазу, 5 771 унификацияланған ҚЖК идеографы. Чироки әріптері.
9.0	2016 маусым	ISO/IEC 10646:2014 қосымша толықтырулар 1 және 2, Адлам, Неуа, Жапон теледидар таңбалары, 74 емоджи мен таңба	135	128 237	Адлам, Байксуки, Маршен, Неуа, Осаге, Таңғұт және 72 емоджи.
10.0	2017 маусым	ISO/IEC 10646:2017 қосымша 56 емоджи таңбасы, 285 хентайгана таңбасы, 3 төртбұрыш Занабазар таңбасы	139	136 755	Занабазар төртбұрыш, Сойомбо, Масарам Гонди, Нушу, Хентайгана, 7 494 унификацияланған ҚЖК идеографы және 56 емоджи.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] ISO/IEC 10646 Әмбебап Кодталған Таңбалар Жиынтығы Халықаралық стандарт-ты [Электрон ресурс]. – URL: <https://www.iso.org/standard/56921.html> (жүгіну уақыты 28.09.2017).

[2] Unicode консорциумының техникалық құжаттары [Электрон ресурс]. – URL: <http://www.unicode.org/reports/index.html#standards> (жүгіну уақыты 28.09.2017).

[3] Unicode кестелері [Электрон ресурс]. – URL: <http://www.unicode.org/charts/> (жүгіну уақыты 28.09.2017).

[4] Қазақ әліпбиінің жана нұсқасы ұсынылды [Электрон ресурс]. – URL: https://kaz.tengrinews.kz/kazakhstan_news/kazak-alpbinn-jana-nuskasyi-usyinyildi-283386/ (жүгіну уақыты 28.09.2017).