Projekt UnicodeBuild

Analiza problemu

Jarosław Kuchta

9.06.2025

# Języki i systemy pisma

Wikipedia definiuje język (ang. language) jako ustrukturalizowany system komunikacji, na który składa się słownik i gramatyka. Język służy przede wszystkim jako środek do komunikacji między ludźmi. Komunikacja ta może być *werbalna* lub *niewerbalna*. Komunikacja werbalna polega na przekazywaniu informacji w formie języka mówionego lub pisanego, z użyciem zrozumiałego dla obu stron komunikacji zbioru słów (*słownika*) i reguł układania słów w sensowne zdania (*gramatyka*). Komunikacja niewerbalna nie używa słów, a informacja może być przekazywana poprzez gesty (tzw. *mowa ciała*), dotyk, mimikę twarzy. Komunikacja niewerbalna może być trudniejsza do wyrażenia i zrozumienia, a także bardziej niejednoznaczna niż komunikacja werbalna. Z drugiej strony komunikacja niewerbalna umożliwia porozumienie między stronami nie znającymi wspólnego języka. Unicode zajmuje się przede wszystkim komunikacją werbalną, pisemną, ale w niektórych przypadkach słowa mogą być zastąpione przez symbole graficzne, w tym reprezentujące np. gesty i mimikę.

Wyróżnia się języki naturalne lub sztuczne. *Języki naturalne* (ang. natural languages) są stosowane wyłącznie do komunikacji między ludźmi. Języki naturalne tworzyły się przez długi czas, w sposób naturalny, a ich reguły były wtórnie określone względem praktycznego użycia. Języki naturalne podlegały (i w dalszym ciągu podlegają) naturalnym zmianom, część *języków wymarła* (ang. extinct languages), tzn. nie są już używane przez żyjących ludzi, a ich poznanie jest możliwe dzięki *artefaktom pisemnym,* czyli inskrypcjom. Niektóre języki wymarłe pozostają nieodkodowane, tzn. znany jest zestaw symboli, ale nieznane jest ich znaczenie. Unicode zajmuje się przede wszystkim językami *żywymi* (tzn. niewymarłymi), ale część standardu dotyczy też języków wymarłych, w tym nieodkodowanych.

W przeciwieństwie do języków naturalnych *języki sztuczne* (ang. artificial languages) zostały opracowane w stosunkowo krótkim czasie, zazwyczaj przez jedną osobę lub niewielką grupę osób, w określonym celu. Języki sztuczne mogą być używane zarówno do komunikacji między ludźmi (np. esperanto, język Braille’a), jak i pomiędzy ludźmi a urządzeniami (np. język APL służący do programowania komputerów) albo też między samymi urządzeniami (bez pośrednictwa ludzi). Unicode zajmuje się językami sztucznymi o ile wymagają one specjalnie zdefiniowanego zestawu znaków.

Komunikacja werbalna może być mówiona lub pisana, tak więc język może być realizowany w *mowie* lub w *piśmie*. Większość języków mówionych ma swoje odpowiedniki w piśmie, ale istnieją też języki, których sposób zapisu nie istnieje lub jest nieznany. Unicode zajmuje się tylko pisemną formą języka, chociaż część systemów pisma stanowi wierne odwzorowanie mowy, a część systemów jest oparta o symbole reprezentujące określone znaczenie, a sposób ich wymowy niekoniecznie jest określony.

*Mowa* (ang. speech) polega na artykułowaniu głosu poprzez użycie ludzkich organów mowy, takich jak krtań, język (ang. tongue), podniebienie miękkie i twarde, dziąsła, wargi. System organów mowy, przez który przechodzi głos nazywa się kanałem głosowym (ang. vocal tract). Mowa może być utrwalona za pomocą środków technicznych (nagranie) do późniejszego odtworzenia. Współcześnie możliwe jest też *generowanie mowy* bez użycia ludzkich organów mowy, lecz środków technicznych dających uformowanie ciągu dźwięków tak, aby przypominały mowę ludzką.

Wypowiedź ustna (ang. oral statement) jest to uporządkowana w czasie sekwencja fonemów, przy czym *fonem* (ang. phoneme) jest abstrakcyjną jednostką mowy, realizowaną w mowie rzeczywistej przez głoski. *Głoska* (ang. phone) jest najmniejszym możliwym elementem wypowiedzi ustnej o stałych cechach fonetycznych (artykulacyjnych i akustycznych). Cechy artykulacyjne reprezentują sposób kształtowania dźwięku, a cechy akustyczne – sposób propagacji dźwięku. Wyróżnia się samogłoski, spółgłoski i półsamogłoski.

*Samogłoski* (ang. vowels) są artykułowane bez zamykania kanału głosowego. Rozkład energii akustycznej w czasie jest jednorodny. W różnych językach używa się różnej liczby samogłosek, od 2 do około 30 (w językach europejskich od 5 do 16). W części systemów pisma samogłoski są zapisywane jako osobne znaki, a w części jako dodatkowe symbole (kropki, kreski etc.) umieszczane ponad, pod lub obok znaków reprezentujących samogłoski.

*Spółgłoski* (ang. consonants) są artykułowane z przymknięciem lub całkowitym zamknięciem kanału głosowego (wyjątek stanowi spółgłoska *h*). Przy artykulacji spółgłoski często następuje faza *plozji* polegająca na nagłym otwarciu kanału głosowego i uwolnieniu przepływu powietrza do jamy nosowej lub jamy ustnej (np. po bokach języka). Większość samogłosek jest artykułowanych przy wydechu powietrza z płuc, ale część spółgłosek (tzw. mlaski lub kliki) powstaje przy nagłym otwarciu zwarcia warg lub języka bez przepływu powietrza z płuc, a niektóre przy krótkim wdechu (cmoki). W różnych językach liczba spółgłosek jest różna (od 6 do 84).

Pojedyncze samogłoski lub zespoły złożone z samogłoski i spółgłosek (jednej lub kilku) tworzą *sylaby* (ang. syllables). Samogłoskę nazywa się wówczas ośrodkiem lub jądrem sylaby (ang. nucleus). W niektórych językach ośrodkiem sylaby może być *spółgłoska sylabiczna* (ang. sonant). Są języki pisane, w których różne sylaby zapisuje się w postaci osobnych znaków. Są to tzw. *sylabariusze* (ang. syllabaries).

Czasami oprócz pojęcia sylaby stosowane jest pojęcie mory. Mora (ang. mora) to najmniejsza jednostka czasu, równa lub krótsza od sylaby, istniejąca w niektórych językach mówionych, w których długość fonetyczna ma znaczenie. Na przykład w języku japońskim nazwa miasta *Osaka* składa się z trzech sylab (*O-sa-ka*), ale czterech mor (*O-o-sa-ka*), ponieważ pierwsza sylaba, jest wymawiana z długą samogłoską.

W ramach poszczególnych języków wyróżnia się dialekty. W językoznawstwie przyjmuje się, że *dialekt* to odmiana języka o cechach fonetycznych, leksykalnych lub gramatycznych, które odróżniają ją od innych odmian tego języka. Granica pomiędzy językiem a dialektem bywa nieostra, tak że nie wiadomo dokładnie, czy modyfikacja słownika i zasad gramatycznych pewnego języka doprowadziła do powstania nowego języka, czy też w dalszym ciągu trzeba mówić o dialekcie języka standardowego. W Unicode kwestia dialektów ogranicza się do języków pisanych. Może być tak, że kilka języków pisanych stosuje ten sam zbiór znaków (np. cyrylicę), ale jego część jest wspólna, a niektóre z języków stosują różne rozłączne podzbiory znaków. Tak jest np. w przypadku języków rosyjskiego, ukraińskiego i białoruskiego. W takich przypadkach Unicode zaznacza to w nazwach opisowych poszczególnych znaków.

Oprócz języków mówionych (ang. spoken languages) istnieją też języki migowe (ang. sign languages) stosowane do komunikacji z ludźmi niesłyszącymi lub niemówiącymi. Istnieje wiele języków migowych, które rozwijały się naturalnie w różnych grupach ludzi, dlatego języki migowe zalicza się do języków naturalnych. Język migowy należy odróżnić od języka ciała (zaliczanego do języków niewerbalnych), gdyż każdy język migowy ma zarówno symbolikę opartą głównie na gestach rąk, jak i reguły stosowania tych gestów. Gesty mogą wyrażać pewne idee (podobnie jak w języku ideograficznym), głoski (jak w alfabecie) lub sylaby. Unicode ma w swoim zakresie blok znaków definiujących gesty (poprzez ułożenie dłoni, palców etc.) stosowane w języku migowym opracowanym w 1974 r. przez Valerie Sutton. Nie definiuje znaczenia gestów, a jedynie ilustruje ich wygląd. Zbiór znaków może służyć do graficznego zapisu różnych języków migowych.

Unicode odnosi się do języków nie bezpośrednio, lecz pośrednio – poprzez tzw. systemy pisma. W językoznawstwie pod pojęciem *systemu pisma* (ang. writing system) rozumie się zbiór symboli, zwany *pismem* (ang. script), służący do utrwalania języka mówionego oraz zestaw reguł opisujących poprawne użycie (i łączenie) tych symboli. Jeden język mówiony może mieć kilka różnych systemów pisma, a jeden system pisma może być stosowany w wielu językach.

Pismo zapisuje język mówiony tak, aby osoba czytająca pismo mogła odtworzyć wypowiedź ustną nawet po tysiącach lat. Do czytania pisma najczęściej stosuje się narządy wzroku, ale są też pisma odczytywane przez dotyk (np. pismo Braille’a). Utrwalenie może przyjmować różne formy. Symbole mogą być napisane na papierze lub pergaminie, wyryte w kamieniu lub na tabliczce glinianej (pismo klinowe), malowane pędzelkiem na tkaninie. Istnieje nawet pismo węzełkowe, które jest utrwalane przez wiązanie węzełków na sznurku, a które przypisuje różne znaczenia różnym rodzajom węzłów i ich kombinacji. W ten sposób symbolem staje się pojedynczy węzeł lub układ węzłów. Pismo Braille’a powstaje poprzez wytłoczenie określonej kombinacji wypukłości w podstawowej siatce 2 x 3 punktów. Pismo może być też zakodowane w postaci cyfrowej w nowoczesnej technologii elektromagnetycznej lub optycznej, tak że potrzeba odpowiednich urządzeń dla odtworzenia wypowiedzi. Tym ostatnim sposobem zapisu Unicode się nie zajmuje. Chociaż zbiory znaków są zakodowane z użyciem 24-bitowego zapisu cyfrowego, to definicje znaków dotyczą graficznej formy zapisu (nawet w przypadku języka migowego lub pisma Braille’a).

Większość punktów kodowych zdefiniowanych w Unicode jest przypisana do jednego z systemów pisma stosowanych w języku naturalnym lub sztucznym. Część jednak reprezentuje symbole graficzne niestosowane w żadnym z nich. Są tam symbole stosowane w matematyce, chemii, muzyce i innych dziedzinach. Trudno nazwać takie zbiory symboli *pismem*. Jeśli te zbiory symboli można skojarzyć z regułami ich stosowania, to taki system pisma będzie nazywany *notacją* (ang. notation). Są jednak takie symbole, dla których nie obowiązują żadne reguły. Takie symbole tworzą po prostu zbiory symboli (ang. symbol sets).

# Klasyfikacja znaków Unicode

Klasyfikacja znaków Unicode, czyli tzw. punktów kodowych (ang. code points), może być dokonywana na podstawie:

* nazwy opisowej znaku,
* kategorii ogólnej znaku,
* bloku Unicode, do którego należy.

Klasyfikacja nie ma charakteru rozłącznego, lecz uzupełniający.

## Jednostki klasyfikacyjne

Każdy znak Unicode, można przypisać do pewnej jednostki organizacyjnej, którą może być:

* *pismo* (ang. script) – zbiór znaków do zapisu języka naturalnego lub sztucznego; z pismem jest skojarzony zestaw reguł określających sposób zapisu i odczytu znaków; zestaw ten są definiowane poza standardem Unicode. Pismo alternatywnie z notacją jest nazywane systemem pisma (ang. writing system).
* *notacja* (ang. notation) – zbiór znaków stosowanych w określonej dziedzinie nauki lub techniki; z notacją jest skojarzony zestaw reguł określających sposób zapisu i odczytu znaków; zestaw ten są definiowane poza standardem Unicode.
* *zbiór symboli* (ang. symbol set) – zbiór znaków stosowanych w różnych dziedzinach; reguł sposobu użycia tych symboli nie ma lub są jedynie zaleceniami.
* *podzbiór symboli* (ang. symbol subset) – wybrany podzbiór ze zbioru symboli.
* *obszar* (ang. area) – taki zbiór znaków, w którym poszczególne znaki nie są definiowane indywidualnie; zamiast tego Unicode jedynie definiuje początek i koniec obszaru.
* *język* (ang. language) – zazwyczaj język, który używa danego znaku jest określany przez pismo, ale w pewnych przypadkach dodatkowo jest opisywany w nazwie znaku.
* *rodzina* (ang. family) – zbiór nadrzędny obejmujący inne jednostki organizacyjne, które mają pewne wspólne cechy; np. wyróżnia się rodzinę pism brahmickich (ang. Brahmi scripts).

Pismo i notację nazywa się *systemami pisma* (ang. writing systems). Klasyfikacja znaków do systemów pisma jest ścisła wynika z danych UCD. Klasyfikacja do pozostałych jednostek jest bardziej uznaniowa i może zależeć od algorytmu tworzonego przez użytkownika.

## Rodzaje pisma

Wyróżnia się kilka rodzajów pisma. Są to:

* *alfabet* (ang. alphabet) – pismo, w którym znaki nazywane są literami, a litery reprezentują głoski (samogłoski lub spółgłoski). Niektóre głoski są reprezentowane przez dwuliterowe sekwencje.
* *sylabariusz* (ang. syllabary) – pismo, w którym znaki reprezentują sylaby (lub mory). Symbol w sylabariuszu zazwyczaj reprezentuje dźwięk spółgłoski, po którym następuje dźwięk samogłoski (jądro). Sylabariusze są stosowane m.in. w języku japońskim i w językach z rodziny Yi ze wschodnich obszarów Chin.
* *alfabet sylabiczny*, *abugida* (ang. abugida) – rodzaj alfabetu, w którym jedna litera reprezentuje najczęściej całą sylabę z domyślną samogłoską. Jeśli w sylabie znajduje się inna samogłoska niż domyślna, to zapisuje się ją za pomocą znaków diakrytycznych. Istnieją również odrębne znaki reprezentujące samogłoski. Alfabety sylabiczne są stosowane m.in. w językach bengalskim i tybetańskim.
* *pismo spółgłoskowe*, *abdżad* (ang. abjad) – rodzaj alfabetu posiadający tylko znaki do zapisu spółgłosek. Samogłoski są domyślane przez czytelnika. Przykłady: pismo arabskie, aramejskie, fenickie, hebrajskie, syryjski.
* *pismo logograficzne* (ang. logographic) – pismo, w którym pojedyncze znaki składają się z dwóch komponentów: komponentu reprezentującego znaczenie (ideę) oraz komponentu reprezentującego wymowę. Pismo chińskie jest przykładem pisma logograficznego.
* *pismo ideograficzne* (ang. ideographic) – pismo (lub raczej notacja), w którym znaki reprezentują pewne abstrakcyjne idee. Ideogramami są np. liczby i symbole matematyczne, symbole waluty, znak ‘&’.
* *pismo piktograficzne* (ang. pictographic) – pismo (lub zbiór symboli), w którym znaki reprezentują pewne znaczenia przez podobieństwo do fizycznych obiektów. Piktogramy były stosowane np. w starożytnych cywilizacjach sumeryjskiej, egipskiej i chińskiej. Z czasem piktogramy ewaluowały w logogramy. Współcześnie symbol koperty jest używany do oznaczenia wychodzącej lub przychodzącej poczty e-mail.
* pismo *hieroglificzne* (ang. hieroglyphs) – starożytne pismo egipskie łączące elementy ideograficzne, logograficzne, sylabiczne i alfabetyczne.
* *pismo logo-sylabiczne* (ang. logo-syllabary) – pismo zawierające dwa osobne zestawy znaków: jeden dla zapisu znaczenia, drugi do zapisu sylabicznego. Przykłady zawierają inskrypcje starożytnych kultur ameryki środkowej.
* *pismo klinowe* (ang. cuneiform) – rodzaj najstarszego pisma logo-sylabicznego utrwalonego na tabliczkach glinianych przez wyciskanie kształtów klinowych za pomocą trzciny.

## Kierunkowość pisma

Alfabety mogą być typu unikameralnego (ang. unicameral), zwanego również jednoszeregowym (ang. unicase). Taki typ alfabetu nie ma oddzielnych liter wielkich i małych. Pismo hebrajskie, arabskie, pisma brahmickie to przykłady alfabetów *unikameralnych*. Z kolei pismo łacińskie, greckie, cyrylica i ormiańskie są bikameralne.

Systemy pisma można podzielić pod względem podziału tekstu na linie i kolejności ich odczytywania. Wyróżnia się:

* pisma poziome i pionowe; w poziomych kolejne znaki tekstu pisane są w liniach poziomych, a pismach pionowych – w liniach poziomych.
* pisma pionowe dzieli się na te, w których pionowe linie tekstu czyta się od strony lewej do prawej lub od prawej do lewej.
* pisma poziome dzieli się na te, w których kolejne znaki w linii czyta się od strony lewej do prawej i te, w których czyta się od prawej do lewej.

Wiele pism z Azji wschodniej może być pisanych w układzie poziomym lub pionowym. Znaki chińskie, koreańskie znaki hangul i japońskie znaki kana mogą być pisane w obu układach, gdyż są symbolami logo-graficznymi lub logo-sylabicznymi na planie kwadratu.

W pismach o liniach układających się poziomo znaki są z reguły pisane i czytane od lewej do prawej (LTR). Są jednak takie (arabskie, hebrajskie), w których znaki są czytane z reguły od strony prawej do lewej (RTL). Ze względu na liczne wyjątki od tej reguły są one nazywane pismami *dwukierunkowymi* (ang. bidirectional scripts).

## Kategorie ogólne znaków

Unicode przypisuje każdy znak do jednej z 30 kategorii ogólnych. Każdą z nich oznacza dwuliterowym skrótem. I tak:

* Do kategorii Lu (Uppercase Letter) zalicza wszystkie wielkie litery alfabetu, a do kategorii Ll (Lowercase Letter) zalicza wszystkie małe litery alfabetu. Oczywiście dotyczy to alfabetów bikameralnych. Do kategorii Lu należy 1858 znaków, a do Ll – 2258 znaków. Wszystkie litery wielkie mają swoje odpowiedniki w literach mały i vice versa.
* Kategoria Lt (Titlecase Letter) obejmuje tylko 31 znaków Unicode. Zgodnie z nazwą należą do niej cztery dwuznaki składające się z wielkie i małej litery alfabetu łacińskiego ‘ǅ’, ‘ǈ’, ‘ǋ’, ‘ǲ’, ale też 27 liter wielkich alfabetu greckiego ze znakami diakrytycznymi, np. „GREEK CAPITAL LETTER ALPHA WITH PROSGEGRAMMENI”.
* Kategoria Lm (Modifier Letter) jest zarezerwowana dla liter (lub innych symboli), które modyfikują cechy fonetyczne innych liter. Zapisywane są zwykle za lub przed modyfikowaną literą). W tym sensie są podobne do znaków diakrytycznych, ale nie łączą się w jednej komórce znakowej z modyfikowanym znakiem, lecz mają swoje własne szerokości. Takie znaki są wykorzystywane w systemach transkrypcji fonetycznej dla określenia odpowiedniej wymowy głoski. Unicode zalicza do tej kategorii 404 znaki.
* Kategoria Lo (Other Letter) jest przeznaczona dla wszystkich innych liter (w tym liter alfabetów unikameralnych), sylab oraz symboli przypominających litery (ang. letterlike symbols) z wyglądu, ale o innym wykorzystaniu (jak np. „FEMININE ORDINAL INDICATOR”). W Unicode jest zdefiniowanych 21518 takich znaków.
* Kategoria Nd (Decimal Number) oznacza cyfry dziesiętne z różnych alfabetów. W sumie jest 760 takich cyfr. Część tych znaków to cyfry dziesiętne alfabetu łacińskiego o zmodyfikowanych kształtach (np. obrysowanych, wytłuszczonych, o kształtach przypominających wyświetlacze cyfrowe.
* Kategoria Nl (Letter Number) obejmuje inne cyfry i liczby, np. z rzymskiego systemu liczbowego, z pisma klinowego. Zdefiniowano 236 takich znaków.
* Kategoria No (Other Number) jest przeznaczona dla innych znaków numerycznych, np. zawiera cyfry dziesiętne w indeksie górnym, znaki ułamków zwykłych jak ‘½’, cyfry połówkowe alfabetu tybetańskiego. Wyodrębniono 915 takich znaków.
* Kategoria Mn (Nonspacing Mark) zawiera znaki diakrytyczne (combining marks) o zerowej szerokości. Modyfikują one cechy fonetyczne liter i sylab. Mogą powodować wymowę nosową, określać samogłoskę w sylabie. Same nie mają szerokości, lecz wyświetlane są wraz z modyfikowanym znakiem (zazwyczaj ponad lub pod). Jest w sumie 2020 znaków diakrytycznych.
* Kategoria Mc (Spacing Mark) zawiera znaki łączone wyświetlane po lub przed modyfikowaną literą lub sylabą, np. znak VISARGA oznaczający w fonologii sanskryckiej bezdźwięczną spółgłoskę szczelinową krtaniową, a wyświetlaną za znakiem jako dwie kropki jedna nad drugą (podobnie do znaku dwukropka). W Unicode jest 4068 takich znaków.
* Kategoria Me (Enclosing Mark) zawiera tylko 13 znaków diakrytycznych obejmujących (enclosing marks). Należą do nich znaki modyfikujące liczebniki w cyrylicy w formie okręgu złożonego z kropek, przecinków lub krzyżyków, wewnątrz których umieszcza się modyfikowany znak liczebnika. Należą „obejmujące” nawiasy, należą koło, romb, trójkąt wewnątrz których umieszcza się inne znaki.
* Kategoria Pc (Connector Punctuation) jest przeznaczona dla znaków interpunkcyjnych łączące, np. krawat (tie) w formie łuku łączącego dwa znaki. Unicode definiuje tylko 10 takich znaków.
* Kategoria Pd (Dash Punctuation) oznacza znak interpunkcyjny będący kreską lub myślnikiem. Wyróżnia się 27 takich znaków należących do różnych alfabetów, o różnych kształtach i przeznaczeniu.
* Kategoria Ps (Open Punctuation) i Pe (Close Punctuation) oznacza znak interpunkcyjny odpowiednio otwierający i zamykający (parę). Należą tu nawiasy okrągłe, kwadratowe i klamrowe, znaki „sufitu” i „podłogi”. Razem 79 i 77 takich znaków. Różnica wynika z różnych kształtów niektórych apostrofów.
* Analogicznie kategoria Pi (Initial Punctuation) i Pf (Final Punctuation) oznacza odpowiednio otwierający i zamykający cudzysłów. Jest odpowiednio 12 i 10 takich znaków.
* Kategoria Po (Other Punctuation) łączy inne znaki interpunkcyjne – nie należące do żadnej innej kategorii znaków interpunkcyjnych. Jest 640 takich znaków zarówno wykorzystywanych wspólnie, jak i należących do różnych alfabetów.
* Kategoria Sm (Math Symbol) grupuje symbole matematyczne – 970 symboli.
* Kategoria Sc (Currency Symbol) zawiera różne znaki waluty – 63 znaki.
* Kategoria Sk (Modifier Symbol) łączy symbole modyfikujące (o takim samym znaczeniu jak Lm), ale niebędące literami – 125 znaków.
* Kategoria So (Other Symbol) to wszystkie inne symbole – razem 7376 znaków.
* Kategoria Zs (Space Separator) oznacza spacje o różnych szerokościach i przeznaczeniu – razem 17 spacji.
* Kategoria Zl (Line Separator) zawiera tylko jeden znak o kodzie \x2028 – separator linii.
* Kategoria Zp (Paragraph Separator) zawiera tylko jeden znak o kodzie \x2029 – separator akapitów.
* Kategoria Cc (Control) zawiera nie znaki widoczne, lecz kody sterujące typu C0 (ASCII) lub C1 – 65 znaków.
* Kategoria Cf (Format) łączy znaki sterujące formatem – kierunkiem pisania dla pism dwukierunkowych, znaki zerowej szerokości (spacje i łączniki) – razem 170 znaków.
* Kategoria Cs (Surrogate) to obszary znaków zastępczych do kodowania znaków 24-bitowych w kodzie 16-bitowym. Zdefiniowanych jest tylko 6 kodów dla oznaczenia początków i końców obszarów.
* Kategoria Co (Private Use) to obszary znaków zastępczych do użytku prywatnego (do dowolnego stosowania). Zdefiniowanych jest tylko 6 kodów dla oznaczenia początków i końców obszarów.
* Kategoria Cn (Unassigned) oznacza nieprzydzielone kody znaków.

## Indeksowane systemy pisma

Jednym ze sposobów klasyfikacji znaku jest sprawdzenie pliku Scripts.txt. W pliku tym określone są zbiory znaków przypisane dla poszczególnych systemów pisma. Ta klasyfikacja może być jednak zawodna. Dla przykładu do pisma Latin przypisany jest zestaw znaków o kodach od \x2160 do \x2182, czyli liczebniki rzymskie od „ROMAN NUMERAL ONE” do „ROMAN NUMERAL TEN THOUSAND”. Również do pisma Latin są przypisane znaki „KELVIN SIGN” i „ANGSTROM SIGN”.

Wśród systemów pisma Unicode jest pozycja o nazwie Common. Do tej klasy należą znaki, które mają zastosowanie w wielu systemach pisma. Jeśli istotne jest sprawdzenie, w których konkretnie systemach pisma stosuje się dany znak, to można to sprawdzić w pliku ScriptExtensions.txt. Można się stamtąd dowiedzieć, że znak diakrytyczny „CARON” jest stosowany w alfabetach opartych na piśmie łacińskim oraz w piśmie Bopomofo. W pliku ScriptExtensions.txt (w odróżnieniu od pliku Scripts.txt) nie stosuje się pełnych nazw pism, lecz ich 4-literowe nazw skrócone zdefiniowane w ISO 15924. W przypadku jakichś wątpliwości można odczytać znaczenie tych 4-literowych skróconych nazw z pliku PropertyValueAliases.txt.

## Bloki Unicode

Plik Blocks.txt definiuje bloki Unicode. Jest w sumie 338 bloków. Każdy blok jest definiowany przez początek i koniec i określa przeznaczenie punktów kodowych mieszczących się w zakresie bloku. Zdecydowana większość bloków przypisuje znaki do systemów pisma – większość do określonego alfabetu. Wówczas wszystkie znaki należące do danego bloku są poprzez plik Scipts.txt przypisane do tego samego pisma. W niektórych blokach znajdują się znaki, które są przypisane do dwóch lub kilku pism.

Są też bloki, które nie są przypisane do żadnego pisma przez Scripts.txt. W tych przypadkach przeznaczenie bloku jest opisane tylko tekstowo. Z pewnych opisów wynika, że bloki zawierają znaki wykorzystywane do pisma migowego, do pisma Braille’a, w dziedzinie matematyki, muzyki, chemii (a raczej alchemii), do transkrypcji fonetycznej Inne bloki zawierają symbole wykorzystywane w pismach ozdobnych, na mapach, ikony i emotki albo wreszcie symbole ogólnego przeznaczenia. Katalog przeznaczenia nie jest zamknięty.

# Analiza nazw opisowych

Podstawę generowania nazw skróconych znaków będzie stanowiła analiza nazw opisowych znaków. Znajdują się one w pliku UnicodeData.txt w kolumnie drugiej, Name.

## Odfiltrowanie nazw specjalnych

Nazwa specjalna rozpoczyna się od znaku ‘<’ i kończy znakiem ‘>’. Jest 65 identycznych nazw „<control>”, które są przypisane do znaków sterujących należących do kategorii Cc. Specyficzne nazwy tych znaków są zebrane w tab. 1. Ze względu na wymaganie zwięzłości wyników trzeba zastosować nazwy skrócone z kolumny drugiej.

Wszystkie inne nazwy specjalne są stosowane do oznaczania początku i końca obszarów kodów, w których poszczególne znaki nie mają indywidualnych definicji. Obszary takie znajdują się na liście bloków Unicode. Znak rozpoczynający dany obszar ma format:

<*nazwa bloku*, First>

a znak kończący:

<*nazwa bloku*, Last>

Ponieważ w specyfikacji Unicode jest 19 takich obszarów, to jest 38 znaków o takich nazwach (tab. 2).

Tab. Nazwy znaków sterujących

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Kod | Skrót | Nazwa | | 00 | NUL | Null | | 01 | SOH | Start of Heading | | 02 | STX | Start of Text | | 03 | ETX | End of Text | | 04 | EOT | End of Transmission | | 05 | ENQ | Enquiry | | 06 | ACK | Acknowledge | | 07 | BEL | Bell | | 08 | BS | Backspace | | 09 | HT | Horizontal Tabulation | | 0A | LF | Line Feed | | 0B | VT | Vertical Tabulation | | 0C | FF | Form Feed | | 0D | CR | Carriage Return | | 0E | SO | Shift Out | | 0F | SI | Shift In | | 10 | DLE | Data Link Escape | | 11 | DC1 | Device Control One | | 12 | DC | Device Control Two | | 13 | DC3 | Device Control Three | | 14 | DC4 | Device Control Four | | 15 | NAK | Negative Acknowledge | | 16 | SYN | Synchronous Idle | | 17 | ETB | End of Transmission Block | | 18 | CAN | Cancel | | 19 | EM | End of medium | | 1A | SUB | Substitute | | 1B | ESC | Escape | | 1C | FS | File Separator | | 1D | GS | Group Separator | | 1E | RS | Record Separator | | 1F | US | Unit Separator | | 7F | DEL | Delete | | |  |  |  | | --- | --- | --- | | Kod | Skrót | Nazwa | | 80 | PAD | Padding Character | | 81 | HOP | High Octet Preset | | 82 | BPH | Break Permitted Here | | 83 | NBH | No Break Here | | 84 | IND | Index | | 85 | NEL | Next Line | | 86 | SSA | Start of Selected Area | | 87 | ESA | End of Selected Area | | 88 | HTS | Character Tabulation Set | | 89 | HTJ | Character Tabulation With Justification | | 8A | VTS | Vertical Tabulation Set | | 8B | PLD | Partial Line Down | | 8C | PLU | Partial Line Up | | 8D | RI | Reverse Index | | 8E | SS2 | Single-Shift 2 | | 8F | SS3 | Single-Shift 3 | | 90 | DCS | Device Control String | | 91 | PU1 | Private Use 1 | | 92 | PU2 | Private Use 2 | | 93 | STS | Set Transmit State | | 94 | CCH | Cancel character | | 95 | MW | Message Waiting | | 96 | SPA | Start of Protected Area | | 97 | EPA | End of Protected Area | | 98 | SOS | Start of String | | 99 | SGCI | Single Graphic Character Introducer | | 9A | SCI | Single Character Introducer | | 9B | CSI | Control Sequence Introducer | | 9C | ST | String Terminator | | 9D | OSC | Operating System Command | | 9E | PM | Privacy Message | | 9F | APC | Application Program Command | |

Tab. . Nazwy specjalne punktów kodowych dla obszarów Unicode

|  |  |
| --- | --- |
| Kod | Nazwa |
| 3400 | <CJK Ideograph Extension A, First> |
| 4DBF | <CJK Ideograph Extension A, Last> |
| 4E00 | <CJK Ideograph, First> |
| 9FFF | <CJK Ideograph, Last> |
| AC00 | <Hangul Syllable, First> |
| D7A3 | <Hangul Syllable, Last> |
| D800 | <Non Private Use High Surrogate, First> |
| DB7F | <Non Private Use High Surrogate, Last> |
| DB80 | <Private Use High Surrogate, First> |
| DBFF | <Private Use High Surrogate, Last> |
| DC00 | <Low Surrogate, First> |
| DFFF | <Low Surrogate, Last> |
| E000 | <Private Use, First> |
| F8FF | <Private Use, Last> |
| 17000 | <Tangut Ideograph, First> |
| 187F7 | <Tangut Ideograph, Last> |
| 18D00 | <Tangut Ideograph Supplement, First> |
| 18D08 | <Tangut Ideograph Supplement, Last> |
| 20000 | <CJK Ideograph Extension B, First> |
| 2A6DF | <CJK Ideograph Extension B, Last> |
| 2A700 | <CJK Ideograph Extension C, First> |
| 2B739 | <CJK Ideograph Extension C, Last> |
| 2B740 | <CJK Ideograph Extension D, First> |
| 2B81D | <CJK Ideograph Extension D, Last> |
| 2B820 | <CJK Ideograph Extension E, First> |
| 2CEA1 | <CJK Ideograph Extension E, Last> |
| 2CEB0 | <CJK Ideograph Extension F, First> |
| 2EBE0 | <CJK Ideograph Extension F, Last> |
| 2EBF0 | <CJK Ideograph Extension I, First> |
| 2EE5D | <CJK Ideograph Extension I, Last> |
| 30000 | <CJK Ideograph Extension G, First> |
| 3134A | <CJK Ideograph Extension G, Last> |
| 31350 | <CJK Ideograph Extension H, First> |
| 323AF | <CJK Ideograph Extension H, Last> |
| F0000 | <Plane 15 Private Use, First> |
| FFFFD | <Plane 15 Private Use, Last> |
| 100000 | <Plane 16 Private Use, First> |
| 10FFFD | <Plane 16 Private Use, Last> |

Wszystkie pozostałe nazwy opisowe znaków podlegają dalszemu przetwarzaniu opisanemu poniżej.

## Rozpoznanie systemów pisma

Bardzo często opisowa nazwa zawiera frazę (słowo lub ciąg słów) kluczową oznaczającą nazwę systemu pisma, np. „LATIN CAPITAL LETTER A”, „CYRILLIC CAPITAL LETTER A”. Zazwyczaj fraza ta jest umieszczona na początku nazwy opisowej znaku (jak w ostatnich przykładach), ale są wyjątki:

1. Nazwa systemu pisma może się w ogóle nie pojawiać w nazwie opisowej znaku. Przykłady: „FEMININE ORDINAL INDICATOR”, „MASCULINE ORDINAL INDICATOR”, „TURNED CAPITAL F”.
2. Nazwa systemu pisma może być poprzedzona modyfikatorem formy znaku, np. „SUPERSCRIPT LATIN SMALL LETTER N”, „PARENTHESIZED LATIN SMALL LETTER A”, „MODIFIER LETTER SMALL GREEK GAMMA”.
3. Słowo, które zazwyczaj oznacza nazwę systemu pisma, może być po prostu przymiotnikiem, np. „WHITE LATIN CROSS” oznacza „biały łaciński krzyż”.

W bazie danych UCD rozpoznano w sumie 163 pisma (ang. scripts) pokazanych w tab. 3. W pierwszej kolumnie podano oficjalną nazwę pisma, w drugiej – rodzaj pisma, w trzeciej – frazę kluczową, a w czwartej 4-literową nazwę skróconą zgodną ze standardem ISO 15924 i opcjonalny alias. W piątej kolumnie pokazano zakresy znaków Unicode, w których znajdują się znaki przypisane do danego pisma. Zakresy te są określone przez jeden lub wiele bloków Unicode.

Tab. . Rozpoznawane pisma Unicode wraz z frazami kluczowymi

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pismo | Rodzaj | Fraza kluczowa | Skrót | RTL | Bikam. |
| Adlam | alphabet | ADLAM\* | adlm | ☑ | ☑ |
| Ahom | abugida | AHOM\* | ahom | ☐ | ☐ |
| Anatolian Hieroglyphs | logographic | ANATOLIAN HIEROGLYPH\* | hluw | ☐ | ☐ |
| Arabic | abjad | ARABIC\* | arab | ☑ | ☐ |
| Armenian | alphabet | ARMENIAN\* | armn | ☐ | ☑ |
| Avestan | alphabet | AVESTAN\* | avst | ☑ | ☐ |
| Balinese | abugida | BALINESE\* | bali | ☐ | ☐ |
| Bamum | pictographic | BAMUM\* | bamu | ☐ | ☐ |
| Bassa Vah | alphabet | BASSA VAH\* | bass | ☐ | ☐ |
| Batak | abugida | BATAK\* | batk | ☐ | ☐ |
| Bengali | abugida | BENGALI\* | beng | ☐ | ☐ |
| Bhaiksuki | abugida | BHAIKSUKI\* | bhks | ☐ | ☐ |
| Bopomofo | semi-syllabary | BOPOMOFO\* | bopo | ☐ | ☐ |
| Brahmi | abugida | BRAHMI\* | brah | ☐ | ☐ |
| Buginese (Lontara) | abugida | BUGINESE\* | bugi | ☐ | ☐ |
| Buhid | abugida | BUHID\* | buhd | ☐ | ☐ |
| Canadian Aboriginal | abugida | CANADIAN SYLLABICS\* | cans | ☐ | ☐ |
| Carian | alphabet | CARIAN\* | cari | ☐ | ☐ |
| Caucasian Albanian | alphabet | CAUCASIAN ALBANIAN\* | aghb | ☐ | ☐ |
| Chakma | abugida | CHAKMA\* | cakm | ☐ | ☐ |
| Cham | abugida | CHAM\* | cham | ☐ | ☐ |
| Cherokee | abugida | CHEROKEE\* | cher | ☐ | ☑ |
| Chorasmian | abjad | CHORASMIAN\* | chrs | ☑ | ☐ |
| CJK Compatibility | logo-syllabary | CJK\* | cjkc | ☐ | ☐ |
| CJK Compatibility Ideographs | ideograph | CJK COMPATIBILITY IDEOGRAPH\* | cjki | ☐ | ☐ |
| CJK Radicals | ideograph | CJK RADICAL\* | cjkr | ☐ | ☐ |
| CJK Strokes | syllabary | CJK STROKE\* | cjks | ☐ | ☐ |
| CJK Unified Ideographs | ideograph | \*CJK UNIFIED IDEOGRAPH\* | cjku | ☐ | ☐ |
| Clear Script | alphabet | MONGOLIAN \* TODO \* | todo | ☑ | ☐ |
| Coptic | alphabet | COPTIC\* | copt | ☐ | ☑ |
| Cuneiform | cuneiform | CUNEIFORM\* | xsux  (cune) | ☐ | ☐ |
| Cypriot | syllabary | CYPRIOT\* | cprt | ☐ | ☐ |
| Cypro-Minoan | syllabary | CYPRO-MINOAN\* | cpmn | ☐ | ☐ |
| Cyrillic | alphabet | CYRILLIC\* | cyrl | ☐ | ☑ |
| Deseret | alphabet | DESERET\* | dsrt | ☐ | ☑ |
| Devanagari | abugida | DEVANAGARI\* | deva | ☐ | ☐ |
| Dives Akuru | abugida | DIVES AKURU\* | diak | ☐ | ☐ |
| Dogra | abugida | DOGRA\* | dogr | ☐ | ☐ |
| Egyptian Hieroglyphs | hieroglyphic | EGYPTIAN HIEROGLYPH\* | egyp | ☐ | ☐ |
| Elbasan | alphabet | ELBASAN\* | elba | ☐ | ☐ |
| Elymaic | abjad | ELYMAIC\* | elym | ☑ | ☐ |
| Ethiopic | abugida | ETHIOPIC\* | ethi | ☐ | ☐ |
| Galik | alphabet | MONGOLIAN \* ALI GALI\* | gali | ☑ | ☐ |
| Garay | alphabet | GARAY\* | gara | ☑ | ☑ |
| Georgian | alphabet | GEORGIAN\* | geor | ☐ | ☑ |
| Glagolitic | alphabet | GLAGOLITIC\* | glag | ☐ | ☑ |
| Gothic | alphabet | GOTHIC\* | goth | ☐ | ☐ |
| Grantha | abugida | GRANTHA\* | gran | ☐ | ☐ |
| Greek | alphabet | GREEK\* | grek  (greek) | ☐ | ☑ |
| Gujarati | abugida | GUJARATI\* | gujr | ☐ | ☐ |
| Gunjala Gondi | abugida | GUNJALA GONDI\* | gong | ☐ | ☐ |
| Gurmukhi | abugida | GURMUKHI\* | guru | ☐ | ☐ |
| Gurung Khema | abugida | GURUNG KHEMA\* | gukh | ☐ | ☐ |
| Hangul | alphabet | HANGUL\* | hang | ☐ | ☐ |
| Hanifi Rohingya | alphabet | HANIFI ROHINGYA\* | rohg | ☑ | ☐ |
| Hanunoo | abugida | HANUNOO\* | hano | ☐ | ☐ |
| Hatran | abjad | HATRAN\* | hatr | ☑ | ☐ |
| Hebrew | abjad | HEBREW\* | hebr | ☑ | ☐ |
| Hentaigana | syllabary | HENTAIGANA\* | hent | ☐ | ☐ |
| Hiragana | syllabary | HIRAGANA\* | hira | ☐ | ☐ |
| Imperial Aramaic | abjad | IMPERIAL ARAMAIC\* | armi | ☑ | ☐ |
| Inscriptional Pahlavi | abjad | INSCRIPTIONAL PAHLAVI\* | phli | ☑ | ☐ |
| Inscriptional Parthian | abjad | INSCRIPTIONAL PARTHIAN\* | prti | ☑ | ☐ |
| Javanese | abugida | JAVANESE\* | java | ☐ | ☐ |
| Kaithi | abugida | KAITHI\* | kthi | ☐ | ☐ |
| Kannada | abugida | KANNADA\* | knda | ☐ | ☐ |
| Katakana | syllabary | KATAKANA\* | kana | ☐ | ☐ |
| Kawi | abugida | KAWI\* | kawi | ☐ | ☐ |
| Kayah Li | alphabet | KAYAH LI\* | kali | ☐ | ☐ |
| Kharoshthi | abugida | KHAROSHTHI\* | khar | ☑ | ☐ |
| Khitan Small | logographic | KHITAN SMALL SCRIPT\* | kits | ☐ | ☐ |
| Khmer | abugida | KHMER\* | khmr | ☐ | ☐ |
| Khojki | abugida | KHOJKI\* | khoj | ☐ | ☐ |
| Khudawadi | abugida | KHUDAWADI\* | sind | ☐ | ☐ |
| Kirat Rai | abugida | KIRAT RAI\* | krai | ☐ | ☐ |
| Lao | abugida | LAO\* | laoo  (lao) | ☐ | ☐ |
| Latin | alphabet | LATIN\* | latn  (latin) | ☐ | ☐ |
| Lepcha | abugida | LEPCHA\* | lepc | ☐ | ☐ |
| Limbu | abugida | LIMBU\* | limb | ☐ | ☐ |
| Linear A | logographic | LINEAR A\* | lina | ☐ | ☐ |
| Linear B | syllabary | LINEAR B\* | linb | ☐ | ☐ |
| Lisu | abugida | LISU\* | lisu | ☐ | ☐ |
| Lycian | alphabet | LYCIAN\* | lyci | ☐ | ☐ |
| Lydian | alphabet | LYDIAN\* | lydi | ☐ | ☐ |
| Mahajani | abugida | MAHAJANI\* | mahj | ☐ | ☐ |
| Makasar | abugida | MAKASAR\* | maka | ☐ | ☐ |
| Malayalam | abugida | MALAYALAM\* | mlym | ☐ | ☐ |
| Mandaic | alphabet | MANDAIC\* | mand | ☑ | ☐ |
| Manichaean | abjad | MANICHAEAN\* | mani | ☑ | ☐ |
| Marchen | abugida | MARCHEN\* | marc | ☐ | ☐ |
| Masaram Gondi | abugida | MASARAM GONDI\* | gonm | ☐ | ☐ |
| Medefaidrin | constructed | MEDEFAIDRIN\* | medf | ☐ | ☑ |
| Meetei Mayek | abugida | MEETEI MAYEK\* | mtei | ☐ | ☐ |
| Mende Kikakui | syllabary | MENDE KIKAKUI\* | mend | ☑ | ☐ |
| Meroitic Cursive | abugida | MEROITIC CURSIVE\* | merc | ☐ | ☐ |
| Meroitic Hieroglyphs | hieroglyphic | MEROITIC HIEROGLYPHIC\* | mero | ☑ | ☐ |
| Miao | abugida | MIAO\* | plrd (miao) | ☐ | ☐ |
| Modi | abugida | MODI\* | modi | ☐ | ☐ |
| Mongolian | alphabet | MONGOLIAN\* | mong | ☐ | ☐ |
| Mro (Mru) | alphabet | MRO\* | mroo (mro) | ☐ | ☐ |
| Multani | abugida | MULTANI\* | mult | ☐ | ☐ |
| Myanmar (Burmese) | abugida | MYANMAR\* | mymr | ☐ | ☐ |
| Nabataean | abjad | NABATAEAN\* | nbat | ☑ | ☐ |
| Nag Mundari | alphabet | NAG MUNDARI\* | nagm | ☐ | ☐ |
| Nandinagari | abugida | NANDINAGARI\* | nand | ☐ | ☐ |
| New Tai Lue | abugida | NEW TAI LUE\* | talu | ☐ | ☐ |
| Newa Prachalit | abugida | NEWA\* | newa | ☐ | ☐ |
| Nko | alphabet | NKO\* | nkoo  (nko) | ☑ | ☐ |
| Nushu | syllabary | NUSHU\* | nshu | ☐ | ☐ |
| Nyiakeng Puachue Hmong | alphabet | NYIAKENG PUACHUE HMONG\* | hmnp | ☐ | ☐ |
| Ogham | alphabet | OGHAM\* | ogam | ☐ | ☐ |
| Ol Chiki | alphabet | OL CHIKI\* | olck | ☐ | ☐ |
| Ol Onal | alphabet | OL ONAL\* | onao | ☐ | ☐ |
| Old Hungarian | alphabet | OLD HUNGARIAN\* | hung | ☐ | ☐ |
| Old Permic | alphabet | OLD PERMIC\* | perm | ☐ | ☐ |
| Old Persian | cuneiform | OLD PERSIAN\* | xpeo | ☐ | ☐ |
| Old Sogdian | abjad | OLD SOGDIAN\* | sogo | ☑ | ☐ |
| Old South Arabian | abjad | OLD SOUTH ARABIAN\* | sarb | ☑ | ☐ |
| Old Turkic | alphabet | OLD TURKIC\* | orkh | ☑ | ☐ |
| Old Uyghur | abjad | OLD UYGHUR\* | ougr | ☑ | ☐ |
| Oriya | abugida | ORIYA\* | orya | ☐ | ☐ |
| Osage | alphabet | OSAGE\* | osge | ☐ | ☑ |
| Osmanya | alphabet | OSMANYA\* | osma | ☐ | ☐ |
| Pahawh Hmong | abugida | PAHAWH HMONG\* | hmng | ☐ | ☐ |
| Palmyrene | alphabet | PALMYRENE\* | palm | ☐ | ☐ |
| Pau Cin Hau | alphabet | PAU CIN HAU\* | pauc | ☐ | ☐ |
| Phags-Pa | alphabet | PHAGS-PA\* | phag | ☐ | ☐ |
| Phoenician | abjad | PHOENICIAN\* | phnx | ☑ | ☐ |
| Psalter Pahlavi | abjad | PSALTER PAHLAVI\* | phlp | ☑ | ☐ |
| Rejang | abugida | REJANG\* | rjng | ☐ | ☐ |
| Runic | alphabet | RUNIC\* | runr | ☐ | ☐ |
| Samaritan | abjad | SAMARITAN\* | samr | ☐ | ☐ |
| Saurashtra | abugida | SAURASHTRA\* | saur | ☐ | ☐ |
| Sharada | abugida | SHARADA\* | shrd | ☐ | ☐ |
| Shavian | alphabet | SHAVIAN\* | shaw | ☐ | ☐ |
| Siddham | abugida | SIDDHAM\* | sidd | ☐ | ☐ |
| Sinhala | abugida | SINHALA\* | sinh | ☐ | ☐ |
| Sogdian | abjad | SOGDIAN\* | sogd | ☑ | ☐ |
| Sora Sompeng | alphabet | SORA SOMPENG\* | sora | ☐ | ☐ |
| Soyombo | abugida | SOYOMBO\* | soyo | ☐ | ☐ |
| Sundanese | abugida | SUNDANESE\* | sund | ☐ | ☐ |
| Sunuwar | alphabet | SUNUWAR\* | sunu | ☐ | ☐ |
| Syloti Nagri | abugida | SYLOTI NAGRI\* | sylo | ☐ | ☐ |
| Syriac | abjad | SYRIAC\* | syrc | ☑ | ☐ |
| Tagalog | abugida | TAGALOG\* | tglg | ☐ | ☐ |
| Tagbanwa | abugida | TAGBANWA\* | tagb | ☐ | ☐ |
| Tai Le | abugida | TAI LE\* | tale | ☐ | ☐ |
| Tai Tham | abugida | TAI THAM\* | lana | ☐ | ☐ |
| Tai Viet | abugida | TAI VIET\* | tavt | ☐ | ☐ |
| Takri | abugida | TAKRI\* | takr | ☐ | ☐ |
| Tamil | abugida | TAMIL\* | taml | ☐ | ☐ |
| Tangsa | alphabet | TANGSA\* | tnsa | ☐ | ☐ |
| Tangut | logographic | TANGUT\* | tang | ☐ | ☐ |
| Telugu | abugida | TELUGU\* | telu | ☐ | ☐ |
| Thaana | abugida | THAANA\* | thaa | ☑ | ☐ |
| Thai | abugida | THAI\* | thai | ☐ | ☐ |
| Tibetan | abugida | TIBETAN\* | tibt | ☐ | ☐ |
| Tifinagh | abjad | TIFINAGH\* | tfng | ☐ | ☐ |
| Tirhuta | abugida | TIRHUTA\* | tirh | ☐ | ☐ |
| Todhri | alphabet | TODHRI\* | todr | ☐ | ☐ |
| Toto | alphabet | TOTO\* | toto | ☐ | ☐ |
| Tulu-Tigalari | abugida | TULU-TIGALARI\* | tutg | ☐ | ☐ |
| Ugaritic | abjad | UGARITIC\* | ugar | ☐ | ☐ |
| Vai | syllabary | VAI\* | vaii | ☐ | ☐ |
| Vithkuqi | alphabet | VITHKUQI\* | vith | ☐ | ☑ |
| Wancho | alphabet | WANCHO\* | wcho | ☐ | ☐ |
| Warang Citi | alphabet | WARANG CITI\* | wara | ☐ | ☑ |
| Yezidi | alphabet | YEZIDI\* | yezi | ☑ | ☐ |
| Yi Syllable | abugida | YI SYLLABLE\* | yiii (yi) | ☐ | ☐ |
| Zanabazar Square | abugida | ZANABAZAR SQUARE\* | zanb | ☐ | ☐ |

Standardowe, 4-literowe nazw skrócone pisma nie zawsze są wygodne w stosowaniu. dlatego w niektórych przypadkach zapewniono aliasy do nich:

* Dla przykładu pismo greckie ma 4-literową nazwę standardową „grek”, a łacińskie „latn”. Dodanie po jeszcze jednej literze ułatwiłoby pisanie tych nazw przez człowieka. Dlatego dodano do nich aliasy (odpowiednio „greek”, i „latin”).
* Standardowa nazwa pisma klinowego „xsux” jest co prawda zgodna z regułami ISO, które nakazują nazywać języki i pisma wymarłe od litery ‘x”, ale mało czytelna. Dlatego do pisma klinowego dodano alias „cune”.
* Pismo o standardowej nazwie „prld” to pismo opracowane przez Sama Pollarda, misjonarza – metodysty dla języka A-Hmao, jednego z rodziny języków Miao stosowanych w południowo-wschodniej Azji. To pismo jest znane pod nazwami Pollard Miao lub Miao, więc dodano do niego alias „miao”.
* Kilka pism ma nazwy standardowe dłuższe niż ich oficjalne nazwy. Są to pisma: Lao, Mro, Nko, Yi. Dla tych pism oprócz standardowych nazw „laoo”, „mroo”, „nkoo” i „yiii” zdefiniowano aliasy, odpowiednio: „lao”, „mro”, „nko” i „yi”.

## Rodzaje pism

Pisma można podzielić na kilka rodzajów (tab. 4). Rodzaje pism nie są jawnie identyfikowane w standardzie Unicode, lecz opisane w Wikipedii. Mogą mieć znaczenie przy parsowaniu opisów punktów kodowych.

Tab. . Rodzaje pism

|  |  |
| --- | --- |
| Rodzaj pisma | Opis |
| alfabet  (*alphabet*) | System pisma oparty na standardowym zestawie liter zapisywanych w celu reprezentowania określonych dźwięków w języku mówionym. |
| sylabariusz (*syllabary*) | system pisma, w którym pojedyncze znaki reprezentują całe sylaby, a nie pojedyncze głoski. |
| spółgłoskowe (*abjad*) | System pisma, w którym reprezentowane są tylko spółgłoski, pozostawiające dźwięki samogłosek do wywnioskowania przez czytelnika. |
| alfabet sylabiczny (*abugida*) | Rodzaj alfabetu, w którym jeden znak reprezentuje zasadniczo całą sylabę wraz z „domyślną” samogłoską. Samogłoska domyślna najczęściej jest dość neutralna, typu szwa, jednak w zależności od języka może mieć różne zabarwienie. Podstawą znaku jest najczęściej spółgłoska, a występujące w sylabie samogłoski inne niż domyślna zapisuje się za pomocą dodania różnorodnych znaków diakrytycznych, choć istnieją również odrębne znaki do zapisu samogłosek. |
| semi-sylabiczne  (*semi-syllabary*) | System pisma, który zachowuje się częściowo jak alfabet, a częściowo jak sylabariusz. |
| sztuczne  (*constructed*) | Pismo opracowane na potrzeby języka, którego fonologia, gramatyka, ortografia i słownictwo nie rozwinęły się w sposób naturalny, ale zostały świadomie wymyślone w jakimś celu |
| klinowe (*cuneiform*) | Logo-sylabiczny system pisma, który był używany do zapisu kilku języków starożytnego Bliskiego Wschodu, nazwany tak od charakterystycznych odcisków w kształcie klina, które tworzą ich znaki. |
| hieroglificzne (*hieroglyphic*) | Starożytny system pisma, który łączy elementy ideograficzne, logograficzne, sylabiczne i alfabetyczne z ponad 1000 różnych znaków. |
| ideograficzne (*ideographic*) | System pisma składający się z symboli, które reprezentują idee lub koncepcje niezależne od konkretnego języka. |
| logograficzne (*logographic*) | System pisma, w którym każda jednostka reprezentuje semantyczny składnik języka, taki jak słowo lub morfem. |
| piktograficzne (*pictographic*) | System pisma składający się z zestawu symboli graficznych, który przekazuje znaczenie poprzez wizualne podobieństwo do obiektów fizycznych. |
| piktofonetyczne  (*pictophonetic*) | System pisma zawierający jednostki składające się z elementów ideograficznych i fonetycznych. |
| mieszane (*mixed*) | System pisma, który ma mieszane cechy pisma logograficznego, sylabariusza i notacji fonetycznej. |

Szczególne rodzaje pism to:

* pisma RTL (Right-To-Left) i dwukierunkowe (bidirectional) – większość pism jest czytanych wierszami od góry do dołu, a znaki w wierszach od lewej do prawej (LTR). Część ma znaki czytane od prawej do lewej (RTL). Spośród nich jest też wiele takich, których zwykły tekst jest czytany od prawej do lewej, ale niektóre wyjątki (np. liczby) od lewej do prawej. Pisma zawierające zarówno tekst RTL, jak i LTR są zwane dwukierunkowymi.
* pisma bikameralne (bicameral) – są to alfabety, które mają inne znaki dla liter wielkich (capital letters) i małych (small letters). Można je rozpoznać po tym, że mają niepuste pola UpperCase i LowerCase w specyfikacji punktów kodowych

Pisma RTL (dwukierunkowe) można rozpoznać wg pola Bidir specyfikacji punktów kodowych. Pole to zawiera jedno-, dwu- lub trójliterowe określenie dwukierunkowości znaku (tab. 5). Do rozpoznania pisma RTL nadają się znaki o kierunkowości R, AL i AN. Jeśli w danym piśmie stosuje się takie znaki, to pismo jest RTL. Jeśli nie – to jest pismem LTR. Kierunkowość L, EN, ES i ET jest stosowana w pismach LTR oraz we fragmentach LTR osadzonych w pismach RTL. Znaki o kierunkowości CS, NSM, BN, WS, ON, B i S są stosowane w obu rodzajach pisma. Kierunkowość LTR, LRO, RLE, RLO, PDF, LRI, RLI, FSI i PDI jest deklarowana tylko dla pojedynczych znaków Unicode sterujących formatowaniem dwukierunkowym (tab. 6).

Tab. . Rodzaje dwukierunkowości znaku

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bidir | Rodzaj znaku | Opis | LTR | RTL |
| L | Left-to-Right | znaki wyświetlane od lewej do prawej | ☑ | ☐ |
| R | Right-to-Left | znaki wyświetlane od prawej do lewej | ☐ | ☑ |
| AL | Arabic Right-to-Left | znaki wyświetlane od prawej do lewej w alfabecie arabskim i podobnych | ☐ | ☑ |
| EN | European Number | cyfry europejskie | ☑ | ☐ |
| ES | European Number Separator | europejskie separatory liczbowe (plus, minus) | ☑ | ☐ |
| ET | European Number Terminator | europejskie terminatory liczbowe (znak hash, symbole waluty, itp.) | ☑ | ☐ |
| AN | Arabic Number | cyfry arabsko-indyjskie | ☐ | ☑ |
| CS | Common Number Separator | wspólne separatory liczbowe | ☑ | ☑ |
| NSM | Nonspacing mark | znaki diakrytyczne bez szerokości | ☑ | ☑ |
| BN | Boundary Neutral | znaki kontrole i formatujące | ☑ | ☑ |
| WS | Whitespace | spacje | ☑ | ☑ |
| ON | Other Neutrals | inne znaki neutralne | ☑ | ☑ |
| B | Paragraph Separator | separator akapitów | ☑ | ☑ |
| S | Segment Separator | separator linii | ☑ | ☑ |
| LRE | Left-to-Right Embedding | osadzone formatowanie od lewej do prawej | ☐ | ☐ |
| LRO | Left-To-Rigth Override | przesłonięte formatowanie od lewej do prawej | ☐ | ☐ |
| RLE | Right-to-Left Embedding | osadzone formatowanie od prawej do lewej | ☐ | ☐ |
| RLO | Right-To-Left Override | przesłonięte formatowanie od prawej do lewej | ☐ | ☐ |
| PDF | Pop Directional Format | przywrócenie poprzedniego formatowania kierunkowego | ☐ | ☐ |
| LRI | Left-to-Right Isolate | izolowane formatowanie od lewej do prawej | ☐ | ☐ |
| RLI | Right-to-Left Isolate | izolowane formatowanie od prawej do lewej | ☐ | ☐ |
| FSI | First Strong Isolate | izolowane formatowanie pierwszego od lewej do prawej | ☐ | ☐ |
| PDI | Pop Directional Isolate | przywrócenie kierunkowości izolowanej | ☐ | ☐ |

## Obsługa dwukierunkowości

W standardzie Unicode znaki są uporządkowane "logicznie", tj. w kolejności, w jakiej mają być interpretowane, a nie "wizualnie" (w kolejności, w której są prezentowane). Wszystkie znaki Unicode dzielone są pod względem obsługi dwukierunkowości na cztery kategorie:

* Znaki o dwukierunkowości silnej – to te, które mają określony kierunek. Przykładami tego typu znaków są: większość znaków alfabetycznych, znaków sylabicznych, ideogramów Han, cyfry spoza języków europejskich lub arabskich oraz znaki interpunkcyjne, które są specyficzne tylko dla tych skryptów.
* Znaki o dwukierunkowości słabej – to te, które mają niejasny kierunek. Przykładami tego typu znaków są cyfry europejskie, cyfry wschodnioarabsko-indyjskie, symbole arytmetyczne i symbole walut.
* Znaki o kierunkowości neutralnej nie pozwalają określić kierunku określić bez kontekstu. Przykłady obejmują separatory akapitów, tabulatory i większość innych białych znaków. Do tej kategorii należą również symbole interpunkcyjne, które są wspólne dla wielu skryptów, takie jak dwukropek, przecinek, kropka i spacja nieprzerywalna.
* Znaki jawnego formatowania, nazywane również "kierunkowymi znakami formatowania", to specjalne sekwencje Unicode, które kierują algorytmem do modyfikowania jego domyślnego zachowania.

Znaki jawnego formatowania dzielą się na "znaczniki", znaki "osadzania", "izolaty" i znaki "przesłonięcia". Ich efekty trwają do momentu wystąpienia separatora akapitu lub znaku "pop".

* Znaczniki - jeśli po "słabym" znaku następuje kolejny "słaby" znak, algorytm przygląda się pierwszemu sąsiedniemu "silnemu" znakowi. Czasami prowadzi to do niezamierzonych błędów wyświetlania. Błędy te są poprawiane lub eliminowane za pomocą znaków "pseudo-silnych". Są dwa takie znaczniki: Znak o kodzie LEFT-TO-RIGHT MARK (LRM) oraz RIGHT-TO-LEFT MARK (RLM). Są one wstawiane w miejscu, w którym zamknięty słaby znak odziedziczy swój kierunek pisania.
* Osadzanie kierunkowych znaków formatowania jest klasyczną metodą Unicode jawnego formatowania i od wersji Unicode 6.3 jest odradzane na rzecz "izolatów". "Osadzanie" sygnalizuje, że fragment tekstu ma być traktowany jako kierunkowo odrębny. Tekst znajdujący się w zakresie osadzanych znaków formatowania nie jest niezależny od otaczającego go tekstu. Ponadto znaki w osadzeniu mogą wpływać na kolejność znaków na zewnątrz. Unicode 6.3 uznał, że osadzenia kierunkowe zwykle mają zbyt silny wpływ na otoczenie i dlatego są niepotrzebnie trudne w użyciu.
* Znaki izolowanego formatowania kierunkowego sygnalizują, że fragment tekstu ma być traktowany jako kierunkowo odizolowany od otoczenia. Począwszy od Unicode 6.3, są to znaki formatowania, które są zachęcane w nowych dokumentach - gdy platformy docelowe są znane z ich obsługi. Te znaki formatujące zostały wprowadzone po tym, jak stało się jasne, że osadzanie kierunkowe zwykle ma zbyt silny wpływ na ich otoczenie, a tym samym jest niepotrzebnie trudne w użyciu. W przeciwieństwie do starszych znaków formatowania kierunkowego "osadzania", "izolaty" nie mają wpływu na kolejność tekstu poza ich zakresem. Izolaty mogą być zagnieżdżane i mogą być umieszczane w osadzeniach i przesłonięciach.
* Znaki przesłonięcia formatowania kierunkowego dopuszczają szczególne przypadki, takie jak numery części (np. aby wymusić pisanie numeru części składającego się z mieszanego języka angielskiego, cyfr i liter hebrajskich od prawej do lewej) i zaleca się ich unikanie, gdy tylko jest to możliwe. Podobnie jak w przypadku innych znaków formatowania kierunkowego, "przesłonięcia" mogą być zagnieżdżone jeden w drugim oraz w osadzaniu i izolatach.

Tab. . Kody znaków sterujących formatowaniem dwukierunkowym

|  |  |  |  |
| --- | --- | --- | --- |
| Kod | Symbol | Nazwa znaku | Znaczenie |
| 202A | LRE | LEFT-TO-RIGHT EMBEDDING | osadzone formatowanie od lewej do prawej |
| 202D | LRO | LEFT-TO-RIGHT OVERRIDE | przesłonięte formatowanie od lewej do prawej |
| 202B | RLE | RIGHT-TO-LEFT EMBEDDING | osadzone formatowanie od prawej do lewej |
| 202E | RLO | RIGHT-TO-LEFT OVERRIDE | przesłonięte formatowanie od prawej do lewej |
| 202C | PDF | POP DIRECTIONAL FORMATTING | przywrócenie poprzedniego formatowania kierunkowego |
| 2066 | LRI | LEFT-TO-RIGHT ISOLATE | izolowane formatowanie od lewej do prawej |
| 2067 | RLI | RIGHT-TO-LEFT ISOLATE | izolowane formatowanie od prawej do lewej |
| 2068 | FSI | FIRST STRONG ISOLATE | izolowane formatowanie pierwszego od lewej do prawej |
| 2069 | PDI | POP DIRECTIONAL ISOLATE | przywrócenie kierunkowości izolowanej |

## Rozpoznawanie języków

Unicode rzadko odwołuje się do języków bezpośrednio, najczęściej robi to pośrednio – przez systemy pisma. W niektórych przypadkach jednak w opisach znaków znajdują się zarówno frazy kluczowe pism, jak i języków. Przypadki te pokazano w tab. 7.

Tab. . Identyfikowane języki w nazwach znaków Unicode

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pismo | Język(i) | Fraza kluczowa języka | Skrót języka | Uwagi |
| Cyrillic | byelorussian +ukrainian | BYELORUSSIAN-UKRAINIAN | beu (beukr) | a) |
| ukrainian | UKRAINIAN | uk (ukr) |
| Canadian Aboriginal | Aivilik | AIVILIK | aiv |  |
| Athabascan | ATHAPASCAN | ath |  |
| Blackfoot | BLACKFOOT | bla |  |
| Carrier | CARRIER | crx |  |
| Naskapi | NASKAPI | nsk |  |
| Nattilik | NATTILIK | ntm |  |
| Nunavik | NUNAVIK | iu (iku) |  |
| Nunavut | NUNAVUT | ikt |  |
| Ojibway | OJIBWAY | oji |  |
| Sayisi | SAYISI | chp |  |
| South-slavey | SOUTH-SLAVEY | xsl |  |
| Bible Cree  (macrolanguage) | BIBLE-CREE | cre | b) |
| Moose Cree | MOOSE-CREE | crm |  |
| Northern Cree | N-CREE | crl |  |
| Thopmpson Cree | TH-CREE | crt |  |
| West Cree | WEST-CREE | crk |  |
| Woods Cree | WOODS-CREE | cwd |  |
| Y Cree | Y-CREE | cry | c) |
| Mongolian | Manchu | MANCHU | mnc | d), e) |
| Xibe | SIBE | sjo |
| Galik | ALI GALI | gal |
| Myanmar | Aiton | AITON | aio |  |
| Shan | SHAN | shn |  |
| Arabic | Sindhi | SINDHI | snd |  |

Uwagi:

1. Unicode nie odnosi się do samego języka białoruskiego, ale do dwóch wariantów cyrylicy (oprócz głównego). Jeden wariant jest stosowany w języku ukraińskim, a drugi w ukraińskim i białoruskim
2. Alfabet sylabiczny rdzennych Kanadyjczyków (ang. Canadian Aboriginal Syllabics) ma kilka wariantów stosowanych w poszczególnych językach autochtonicznych Kanady. Wariant Bible Cree odnosi się do makrojęzyka Cree.
3. Nie jest jasne, co oznacza wariant Y-Cree. Nie istnieje oficjalny język o nazwie Y-Cree w standardzie ISO 639-3. Możliwe, że tak jest przywoływany jeden z dialektów języka Cree. Skrót nazwy „cry” został utworzony sztucznie.
4. Oprócz wymienionych tutaj języków mongolskich w standardzie Unicode zdefiniowane są też znaki dla pisma Clear Script i Galik. Wszystkie one mają opisy zaczynające się od „MONGOLIAN”.
5. Niektóre znaki Unicode są wykorzystywane zarówno w języku Manchu, jak i Galik. Składnia opisu jest podobna do „MONGOLIAN LETTER MANCHU ALI GALI \*”. Litery, które należą tylko do języka Galik mają składnię podobną do „MONGOLIAN LETTER ALI GALI \*”, a do Manchu – „MONGOLIAN LETTER MANCHU \*”. Są jeszcze znaki nieliterowe języka Manchu o składni „MONGOLIAN MANCHU \*”.

## Rozpoznawanie systemów notacji

Dalsza analiza przeznaczenia znaków i bloków Unicode umożliwiła rozpoznanie szeregu *notacji,* czyli systemów pisma niepowiązanych (przynajmniej bezpośrednio) z żadnym konkretnym językiem naturalnym. Są one wymienione w tab. 8. Wyróżnia się wśród nich notacje typu:

* liczbowego (ang. numeric) – do zapisu liczb w różnych systemach
* matematycznego (ang. mathematical) – do zapisu wzorów i wyrażeń matematycznych,
* muzycznego (ang. musical) – do zapisu muzyki,
* metrycznego (ang. metrical) – do oznaczania rytmu (długości sylab) w poezji (zwłaszcza łacińskiej i greckiej),
* fonetycznego (ang. phonetic) – do zapisu wskazówek dotyczących wymowy,
* stenograficznego (ang. stenographic) – do zapisywania mówionego tekstu w sposób skrócony,
* gestykulacyjnego (ang. gestural) – do zapisywania gestów języka migowego.

Tab. . Notacje Unicode wyróżnione w UCD

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Notacja | Rodzaj | Fraza kluczowa | Skrót | RTL | Uwagi |
| Aegean Numbers | numeric | AEGEAN | AEGN | ☐ | a) |
| APL | symbolic | APL FUNCTIONAL SYMBOL | APL |  | b) |
| Arabic-Indic Numeric Notation | numeric | ARABIC-INDIC | ARIN | ☑ | c) |
| Balinese Musical Symbols | musical | BALINESE MUSICAL SYMBOL | BLMN |  |  |
| Byzantine Musical Symbols | musical | BYZANTINE MUSICAL SYMBOL | BZMN |  |  |
| Coptic Epact Numbers | numeric | COPTIC EPACT | COPN | ☐ |  |
| Counting Rod Numerals | numeric | COUNTING ROD | CNTR | ☐ |  |
| Duployan Shorthand | stenographic | DUPLOYAN | DUPL |  |  |
| Extended Arabic-Indic Numeric Notation | numeric | EXTENDED ARABIC-INDIC | AREN | ☐ | c) |
| Garay Digits | numeric | GARAY DIGIT | GARA | ☑ |  |
| Greek Acrophonic Numbers | numeric | GREEK ACROPHONIC | GRAN | ☐ |  |
| Greek Ancient Instrumental Notation | musical | GREEK INSTRUMENTAL NOTATION | GRIN |  |  |
| Greek Ancient Vocal Notation | musical | GREEK VOCAL NOTATION | GRVN |  |  |
| Greek Metrical System | metrical | METRICAL | GRMS |  |  |
| Greek Musical Symbols | musical | GREEK MUSICAL | GRMN |  |  |
| Hanifi Rohingya Digits | numeric | HANIFI ROHINGYA DIGIT | RIGH | ☑ |  |
| Indic Siyaq Numbers | numeric | INDIC SIYAQ | ISN | ☐ | d) |
| Kaktovik Numerals | numeric | KAKTOVIK NUMERAL | KTVN | ☐ |  |
| Kanbun | phonetic | IDEOGRAPHIC ANNOTATION | KANB |  |  |
| Kievan Musical Symbols | musical | MUSICAL SYMBOL KIEVAN | KVMN |  |  |
| Mathematical Notation | mathematical | MATHEMATICAL | MATH |  | e) |
| Mayan Numerals | numeric | MAYAN NUMERAL | MAYN | ☐ |  |
| Modern Arabic Mathematical | mathematical | ARABIC MATHEMATICAL | AMATH |  |  |
| Musical Notation | musical | MUSICAL SYMBOL | MMN |  |  |
| Ottoman Siyaq Numbers | numeric | OTTOMAN SIYAQ | SIYN |  | f) |
| Persian Musical Symbols | musical | MUSICAL SYMBOL \* | PSMN |  |  |
| Phaistos Disk | symbolic | PHAISTOS DISK | PHSD |  |  |
| Roman Numbers | numeric | ROMAN NUMERAL | ROMN | ☐ |  |
| Rumi Numeral Symbols | numeric | RUMI | RUMI | ☑ |  |
| Signwriting | gesture | SIGNWRITING | SWRT |  |  |
| Suzhou Numbers | numeric | HANGZHOU | SZHN (HZHN) | ☐ |  |
| Z Notation | mathematical | Z NOTATION | ZNOT |  |  |
| Znamenny Chant | musical | ZNAMENNY | ZNMN |  |  |

Uwagi:

1. Notacja Aegean Numbers zawiera nie tylko liczby, lecz też symbole jednostek wagi i innych miar.
2. APL jest notacją używaną w języku opracowanym w 1960 r. przez Kennetha E. Iversona, stosowanym przez IBM jako dodatek do języka Basic.
3. Cyfry wschodnioarabskie arabskie, zwane również cyframi indoarabskimi lub cyframi arabsko-indyjskimi, są symbolami używanymi do reprezentowania cyfr w połączeniu z alfabetem arabskim w krajach Maszrik (wschód świata arabskiego), Półwyspu Arabskiego i jego odmianie w innych krajach, które używają cyfr perskich na płaskowyżu irańskim i w Azji. W Unicode są dwa zakresy cyfr pochodzących z tego zbioru symboli o wartościach od 0 do 9 (te same wartości, ale inne kształty).
4. Indic Siyaq Numbers to podzbiór [pisma arabskiego](https://en.wikipedia.org/wiki/Arabic_script), który był używany w księgowości w [Indiach](https://en.wikipedia.org/wiki/India) pod rządami [Mogołów](https://en.wikipedia.org/wiki/Mughal_Empire) od XVII wieku do połowy XX wieku. Obejmuje on 69 symboli (nie tylko cyfr).
5. Notacja matematyczna (ang. Mathematical Notation) zawiera cyfry dziesiętne symbole (operatory) z kategorii Sm i jest też używana dla liter alfabetu łacińskiego, greckiego i hebrajskiego do zaznaczenia, że są one stosowane w zapisie matematycznym, zazwyczaj do oznaczania zmiennych. Ten zbiór znaków zawiera też podzbiory liter łacińskich pisanych różnymi krojami pisma.
6. Ottoman Siyaq Numbers to podzbiór pisma arabskiego używany do księgowania w dokumentach z czasów Imperium Osmańskiego.
7. Nowoczesna arabska notacja matematyczna (ang. Modern Arabic Mathematical notation) to notacja matematyczna oparta na piśmie arabskim, używana zwłaszcza na poziomie edukacji przed-uniwersyteckiej. Jej forma w większości wywodzi się notacji zachodniej, ale jest zapisywana od prawej do lewej zgodnie z normalnym kierunkiem pisma arabskiego. Inne różnice obejmują zastąpienie liter alfabetu greckiego i łacińskiego dla symboli literami arabskimi oraz używanie nazw arabskich dla funkcji i relacji.
8. Counting Rod Numbers to blok Unicode zawierający tradycyjne chińskie symbole prętów liczących, których matematycy używali do obliczeń w starożytnych Chinach, Japonii, Korei i Wietnamie. Blok zawiera również pięć ideograficznych znaczników liczbowych, które są szeroko stosowane w Azji Wschodniej.
9. System liczbowy Suzhou (oznaczany w Unicode przez frazę kluczową HANGZHOU) to zbiór 9 cyfr dziesiętnych (od 1 do 9) używanych na chińskich targowiskach. Jest jedyną zachowaną odmianą system prętowego. W większości wyparty przez system cyfr hinduskich (indyjsko-arabskich).
10. Roman Numbers to zbiór symbol pochodzących z alfabetu łacińskiego używanych w liczbowej notacji rzymskiej. Mimo iż symbole te wyglądają jak litery łacińskie (z wyjątkiem tylko jednego), to mają osobne kody Unicode.
11. Rumi Numeral Symbols to blok Unicode zawierający znaki numeryczne używane w Fezie, Maroku i innych miejscach w Afryce Północnej i na Półwyspie Iberyjskim między X a XVII wiekiem.
12. System liczbowy Majów był systemem do reprezentowania liczb i dat kalendarzowych w cywilizacji Majów. Był to system liczbowy pozycyjny dwudziestkowy (podstawa 20). Cyfry składają się z trzech symboli: zera (muszli), jedności (kropki) i piątki (kreski). Na przykład trzynaście jest zapisywane jako trzy kropki w poziomym rzędzie nad dwiema poziomymi kreskami; czasami jest również zapisywane jako trzy pionowe kropki na lewo od dwóch pionowych kresek. Za pomocą tych trzech symboli można zapisać każdą z dwudziestu cyfr dwudziestkowych.
13. System liczbowy Kaktovik to system o podstawie 20 stworzony przez alaskańskich Inuitów, używany też w językach Aleutów i ludów Yupik.
14. Coptic Epact Numbers to specjalny system znaków numerycznych używany w niektórych regionach Egiptu w czasach późnego antyku, głównie w języku koptyjskim. Zamiast używać liter alfabetu koptyjskiego do zapisu liczb (jak to robiono w wielu kulturach, np. w Grecji czy Rzymie), system ten wykorzystywał oddzielne znaki dla każdej cyfry, dziesiątki i setki
15. Greckie numeryczne symbole akrofoniczne reprezentują wartości liczbowe używane w starożytnym greckim systemie numeracji, który opierał się na zasadach akrofonicznych – co oznacza, że symbole pochodzą od pierwszej litery odpowiadającego greckiemu słowu dla liczby. Były one używane w różnych wariantach, które w nazwie opisowej znaku Unicode rozróżniane są przez dodatkowe słowa kluczowe (tab. 9). Część symboli wyrażała jednostki monetarne i inne jednostki miary rozróżniane przez słowa kluczowe pokazane w tab. 10.

Tab. . Warianty greckich symboli akrofonicznych

|  |  |  |
| --- | --- | --- |
| Słowo kluczowe | Wariant | Opis |
| ATTIC | attycki | odnosi się do Attyki – starożytnego regionu Grecji obejmującego miasto Ateny i okolice |
| HERAEUM | herański | odnosi się do świątyni bogini Hery |
| THESPIAN | tespiański | związany teatrem lub dramatem (od imienia Thespis, legendarnego aktora z VI wieku p.n.e.) |
| CYRENAIC | cyrenajski | odnosi się do regionie Cyreny, starożytnej greckiej kolonii w północnej Afryce |
| EPIDAUREAN | epidaurejski | odnosi się do Epidauru – starożytnego greckiego miasta w Argolidzie, znanego przede wszystkim jako miejsce kultu boga Asklepiosa, patrona medycyny |
| TROEZENIAN | trojzeński | odnosi się do mieszkańca lub czegoś związanego ze starożytnym miastem Troizena. położonego w regionie Argolidy na Peloponezie w Grecji. |
| CARYSTIAN | karystyjski | odnosi się do Carystos, starożytnego miasta-państwa położonego na południowym wybrzeżu wyspy Eubea w Grecji. |
| NAXIAN | naksoski | odnosi się do wyspy Naksos – największej z Cyklad na Morzu Egejskim |
| DELPHIC | delficki | odnosi się do starożytnego greckiego sanktuarium w Delfach lub do wyroczni delfickiej, czyli kapłanki boga Apollina (Pytii), która udzielała odpowiedzi na pytania pielgrzymów |
| STRATIAN | stratyjski | odnosi się do regionu Stratos, miasta w starożytnej Achai |

Tab. . Jednostki monetarne i inne jednostki miary starożytnej Grecji

|  |  |  |
| --- | --- | --- |
| Słowo kluczowe | Jednostka | Opis |
| DRACHMA(S) | drachma | Starożytna waluta emitowana przez wiele miast-państw w okresie dziesięciu stuleci, od okresu archaicznego, przez okres klasyczny, okres hellenistyczny aż do okresu rzymskiego. Moneta, zwykle wykonana ze srebra, a czasami ze złota, miała swoje początki w systemie barterowym, w którym drachmę odnoszono do garści drewnianych strzał. Drachma była unikalna dla każdego miasta-państwa, które ją biło, i czasami były w obiegu w całym basenie Morza Śródziemnego. Drachma ateńska była uważana za najsilniejszą i stała się najpopularniejszym środkiem płatniczym. |
| QUARTER | jedna czwarta | zapewne jedna czwarta część drachmy |
| TALENT(S) | talent | duża jednostka wagowo-walutową, używana m.in. w kontekście danin, podatków i skarbu państwa. Jeden talent odpowiadał około 26–36 kg srebra. |
| STATER(S) | stater | starożytna srebrna moneta o dużej wartości używana w różnych regionach Grecji |
| MNA(S) | mina | jednostka wagi i waluty. Jedna mina odpowiadała około 100 drachmom lub prawie 0,5 kg srebra. |
| PLETHRON | plethron | starożytna grecka jednostka miary, używana zarówno jako jednostka długości, jak i powierzchni. Jako jednostka długości odpowiadała około 100 stóp greckich, czyli mniej więcej 30 metrów. Jako jednostka powierzchni oznaczała kwadrat o bokach jednego plethronu, czyli około 900 m². |

1. Grecka starożytna notacja muzyczna składa się dwóch osobnych systemów: notacji wokalnej i notacji instrumentalnej. Symbole z tych systemów nie są nazwane, lecz numerowane. Oba systemy mają po 54 symboli.
2. Grecki system metryczny był używany w analizie metrycznej poezji starogreckiej, w której metrum nie opierało się na akcencie, lecz na długości sylab. Każda sylaba mogła być krótka (1 mora) lub długa (2 mory).
3. Kanbun to klasyczny system zapisu i czytania chińskich tekstów literackich w Japonii, stosowany od okresu Nara (VIII w.) aż do XX wieku. Choć teksty były pisane po chińsku, Japończycy opracowali specjalny sposób ich czytania i interpretacji, dzięki któremu przy zachowaniu oryginalnego brzmienia chińskiego można było rozumieć ja na sposób japoński.
4. Unicode obsługuje kilka notacje muzycznych: balijską, bizantyjską, kijowską, perską, współczesną notację muzyczną oraz greckie symbole długości muzycznej – potrójnej (triseme), poczwórnej (tetraseme) i pięciokrotnej (pentaseme) i pauzy (leimma). Greckie symbole długości muzycznej są znakami diakrytycznymi (są kombinowane z innymi symbolami). Perska notacja muzyczna ma tylko dwa symbole, które można rozpoznać przez początkową frazę MUSICAL SYMBOL, która jednak nie rozwija się w żadną inną frazę notacji muzycznych.
5. Notacja Znamenny Chant odnosi się do tradycji śpiewu stosowanej w Rosyjskim Kościele Prawosławnym. Pieśni w tej notacji nie są zapisywane za pomocą nut, lecz specjalnych znaków zwanych proporcami (*znamëna*) lub hakami (*kryuki*).
6. System zapisu znaków (ang. signwriting) opisuje kształty, lecz nie opisuje znaczenia gestów. Dla przykładu SIGNWRITING HAND-FIST INDEX oznacza znak dłoni – pięść z wysuniętym palcem wskazującym.
7. Notacja Duployan Shorthand zawiera symbole stosowane w różnych systemach stenografii, w tym francuskiego Duployéana, pisma Chinook, stenograficznego zapisu rumuńskiego oraz angielskich stenograficznych systemów Sloan- Duployéana, Pernina i Perraulta.

Nie wszystkie notacje zostały w ten sposób wyróżnione. W szczególności notacje fonetyczne IPA (ang. International Phonetic Notation) i UPA (ang. Uralic Phonetic Notation) są oparte odpowiednio o alfabet łaciński i fiński (pismo łacińskie z rozszerzeniami) i nie są odróżniane przez żadną frazę kluczową.

Notacje numeryczne mogą być ukierunkowane LTR lub RTL.

## Typy znaków

Wiele znaków Unicode ma w opisie słowo lub frazę kluczową oznaczającą typ znaku.

|  |  |
| --- | --- |
| Fraza kluczowa | Znaczenie |
| ACCENT | Akcent |
| ARROW | Strzałka |
| ARROWHEAD | Grot strzałki |
| BARB | Asymetryczny grot harpuna |
| CHARACTER | Znak |
| COMBINING | Znak diakrytyczny bez szerokości |
| CONSONANT | Spółgłoska |
| DENOMINATOR | Mianownik |
| DIGIT | Cyfra |
| FRACTION | Ułamek |
| FULL STOP | Kropka (na zakończenie zdania) |
| HARPOON | Harpun (strzałka z grotem asymetrycznym) |
| HIEROGLYPH | Hieroglif |
| IDEOGRAPH | Ideogram |
| LETTER | Litera |
| LIGATURE | Ligatura |
| MARK | Marker |
| MODIFIER | Znak modyfikujący |
| NUMBER | Liczba |
| NUMERAL | Liczebnik |
| NUMERATOR | Licznik |
| POINT | Punkt (dotyczy języka hebrajskiego) |
| PUNCTUATION | Znak interpunkcyjny |
| SEPARATOR | Separator |
| SIGN | Znak |
| SYLLABLE | Sylaba |
| SYMBOL | Symbol |
| TAG | Znacznik |
| TAIL | Zakończenie (zwłaszcza strzałki) |
| TONE | Intonacja |
| VARIATION SELECTOR | Selektor odmiany |
| VOWEL | Samogłoska |

## Rozpoznawanie znaków diakrytycznych

Unicode wyróżnia dwa rodzaje znaków diakrytycznych:

* symbol (lub litery) „kombinatoryczne” (ang. combining), czyli takie, które nie mają własnej szerokości i są dodawane ponad lub pod modyfikowanym znakiem (kategoria Mn), a czasami go obejmują (kategoria Me).
* litery i symbole modyfikujące (ang. modifying letters), czyli takie, które mają własną szerokość i są dodawane po lub przed modyfikowanym znakiem (kategorie Lm i Sk).

Oprócz tego niektóre symbole kombinatoryczne są zdefiniowane jako osobne punkty kodowe z własną szerokością tak, że mogą być używane jako osobne znaki.

Tab. . Identyfikowane języki w nazwach znaków Unicode

|  |  |  |  |
| --- | --- | --- | --- |
| Pismo | Fraza kluczowa | Kat. znaków | Uwagi |
| Common | MODIFIER LETTER\* | Sk, Lm |  |
| Common | MODIFIER\* | Sk |  |
| Common | COMBINING\* | Mn, Me |  |
| Latin | COMBINING LATIN\* | Mn |  |
| Cyrillic | COMBINING CYRILLIC\* | Mn, Me |  |
| Cyrillic | MODIFIER LETTER CYRILLIC\* | Lm |  |
| Armenian | ARMENIAN MODIFIER LETTER\* | Sk |  |
| Nko | NKO COMBINING\* | Mn |  |
| Samaritan | SAMARITAN MODIFIER LETTER\* | Lm |  |
| Telugu | TELUGU SIGN COMBINING\* | Mn |  |
| Kannada | KANNADA SIGN COMBINING\* | Mn |  |
| Malayam | MALAYALAM SIGN COMBINING\* | Mn |  |
| Georgian | MODIFIER LETTER GEORGIAN\* | Sk |  |
| Ethiopic | ETHIOPIC COMBINING\* | Mn |  |
| Tai Tham | TAI THAM COMBINING\* | Mn |  |
| Balinese | BALINESE MUSICAL SYMBOL COMBINING\* | Mn |  |
| Coptic | COPTIC COMBINING\* | Mn |  |
| Tiffinagh | TIFFINAGH MODIFIED\* | Lm |  |
| Katakana, Hiragana | COMBINING KATAKANA-HIRAGAN\* | Mn |  |
| Bamum | BAMUM COMBINING\* | Mn |  |
| Han | MODIFIER LETTER CHINESE\* | Sk |  |
| Devanagari | COMBINING DEVANAGARI\* | Mn |  |
| Myanmar | MYANMAR MODIFIER LETTER\* | Lm |  |
| Phaistos Disc | PHAISTOS DISC SIGN COMBINING\* | Mn |  |
| Old Permic | COMBINING OLD PERMIC\* | Mn |  |
| Garay | GARAY COMBINING\* | Mn |  |
| Yezidi | YEZIDI COMBINING\* | Mn |  |
| Arabic | ARABIC COMBINING\* | Mn |  |
| Sogdian | SOGDIAN COMBINING\* | Mn |  |
| Old Uyghur | OLD UYGHUR COMBINING\* | Mn |  |
| Sharada | SHARADA VOWEL MODIFIER MARK\* | Mn |  |
| Grantha | GRANTHA SIGN COMBINING\* | Mn |  |
| Grantha | COMBINING GRANTHA\* | Mn |  |
| Bindu | COMBINING BINDU\* | Mn |  |
| Egyptian Hieroglyphs | EGYPTIAN HIEROGLYPH MODIFIER\* | Mn |  |
| Bassa Vah | BASSA VAH COMBINING\* | Mn |  |
| Miao | MIAO SIGN CONSONANT MODIFIER BAR\* | Mn |  |
| Znamenny Chant | ZNAMENNY COMBINING\* | Mn |  |
| Znamenny Chant | ZNAMENNY PRIZNAK MODIFIER\* | Mn |  |
| Musical Notation | MUSICAL SYMBOL COMBINING\* | Mc |  |
| Greek Ancient Instrumental Notation | COMBINING GREEK MUSICAL\* | Mn |  |
| Signwriting | SIGNWRITING \* MODIFIER\* | Mn | a) |
| Glagolitic | COMBINING GLAGOLITIC\* | Mn |  |
| Mende Kikakui | MENDE KIKAKUI COMBINING NUMBER\* | Mn |  |
| Adlam | ADLAM \* MODIFIER\* | Mn | b) |
| Emoji | EMOJI MODIFIER FITZPATRICK\* | Sk |  |

Uwagi:

1. Modyfikatory pisma migowego mają określenie typu pomiędzy SIGNWRITING a MODIFIER. Może to być FILL albo ROTATION.
2. Modyfikatory pisma ADLAM mają określenie typu pomiędzy ADLAM a MODIFIER. Może to być CONSONANT albo GEMINATE CONSONANT.

## Rozpoznawanie wielkości liter

W Unicode w alfabetach bikameralnych litery wielkie są oznaczane jako „CAPITAL LETTER”, a litery małe jako „SMALL LETTER”. Ponieważ jednak słowo „SMALL” może oznaczać zwyczajne pomniejszenie symbolu, więc bardziej niezawodne jest rozpoznanie wielkości litery wg kategorii ogólnych (Lu, Ll i Lt).