# IONCHÓDÚ TÉACS AR RÍOMHAIRÍ

*Michal Boleslav Měchura*

An úsáid is coitianta a bhaintear as ríomhairí na laethanta seo ná láimhseáil téacs, idir phróiseáil focal, an ríomhphost, agus téacs a fhoilsiú nó a léamh ar an Ghréasán Domhanda. Go deimhin, ba é próiseáil focal a chéadspreag an pobal i gcoitinne chun glacadh le ríomhairí mar chuid den saol. Nuair a tháinig ríomhairí ar an fhód don chéad uair, áfach, ba chun uimhreacha a phróiseáil a cruthaíodh iad seachas téacs. Nuair a tháinig sé chun cinn go bhfuil gá le téacs a thaifeadadh i gcuimhne an ríomhaire freisin, ba i bhfoirm uimhreacha a rinneadh gach litir na haibítre a thaifeadadh. Glaoitear ionchódú carachtar ar an phrionsabal seo.

Níor smaoineamh nua é, dar ndóigh. Go stairiúil, cruthaíodh go leor córas chun litreacha a thaifeadadh mar rud éigin eile, mar shampla cód Morse agus an tOgham. Ar ríomhairí, déantar uimhir faoi leith a shannadh do gach litir na haibítre – san ionchódú ASCII, cuir i gcás, freagraíonn an uimhir `65` don litir `A`. Ach is iomaí claochlú a tháinig ar an phrionsabal lom simplí seo ó aimsir ASCII. Áit nach raibh ríomhairí in ann ach téacs in aibítir an Bhéarla a thaifeadadh fiche bliain ó shin, tá córais curtha i dtreo anois atá in inmhe téacs in aon teanga ar domhan a láimhseáil i gceart. Is fiú súil siar a chaitheamh ar fhorbairt stairiúil an réimse seo, mar sin, le go dtuigfimis cén dóigh ar éirigh leis an teicneolaíocht freastal ar iliomad teangacha an domhain, fiú teangacha le scríbhneoireacht chasta ar nós na Sínise.

## *An saol faoi ghabháil ag na seacht ngiotán*

Ceann de na chéad scéimeanna ionchódaithe a cruthaíodh nó ASCII (giorrúchán a sheasann do *American Standard Code for Information Interchange*, is é “ascaí” an fhuaim a chuirtear air de ghnáth). Sa bhliain `1963` a seoladh an córas sin agus bhí glactha leis ar fud an tionscail sul i bhfad mar chaighdeán taifeadta agus malartaithe téacs – cé go raibh ar a laghad mórchomhlacht amháin ann, IBM, nár ghlac le ASCII riamh agus a d’fhorbair a chóras féin, EBCDIC.

Luath-thréimhse na ríomhaireachta a bhí ann sna naoi déag seascaidí. Ba é seacht ngiotán gnáthfhad an bhirt ag an chuid ba mhó de phróiseálaithe na linne sin, agus chloígh ASCII – chomh maith le córais eile mar é – leis an uasteorainn sin. Go bunúsach, is tábla dhá cholún é gach ionchódú, tábla a liostaíonn na litreacha ar fad is féidir a thaifeadadh, agus na huimhreacha a fhreagraíonn dóibh. Ó tharla gurb é `127` an uimhir is airde is féidir a léiriú taobh istigh de spás seacht ngiotán ($1111111_2 = 127_{10}$), níl ach `128` litir sa tábla ASCII (ó `0` go `127`). Cuireann sé seo

srianta tromchúiseacha ar an líon teangacha ar féidir le ASCII freastal orthu. I ndáiríre, ní fhreastalaíonn ASCII ach ar an Bhéarla agus ar chorrtheanga eile nach n-úsáideann ach na litreacha A – Z, gan síntí fada nó maisiúcháin eile.

Seo iad na carachtair a fhaightear ar thábla ASCII: na ceannlitreacha A – Z, na litreacha beaga a – z, na huimhreáin 0 – 9, an spás, roinnt siombailí poncaíochta (an lánstad, an chamóg, agus araile) agus cúpla siombail eile a bhí in úsáid go forleathan i saol na ríomhaireachta ag an am sin, leithéidí #, $, @. Anuas orthu seo tá 33 carachtar rialúcháin in ASCII, is é sin, carachtair a thugann treoracha áirithe don ríomhaire: an cúlspás agus aisfhilleadh carráiste, mar shampla. I measc na gcarachtar rialúcháin tá go leor siombailí nach bhfuil ciall leo níos mó agus cuid eile nach raibh ciall leo riamh mar ní carachtair iad ar chor ar bith, mar shampla an clog (treoir don ríomhaire bíp a ligint) agus fotha foirme (treoir don phrintéir bogadh go dtí an chéad leathanach eile). Is meascán mearaí de charachtair agus treoracha nach carachtair iad i ndáiríre é ASCII, agus ar bhreathnú siar dúinn anois, is féidir a rá gur botún a bhí ann na carachtair rialúcháin go léir seo a liostáil ar ASCII.

Ach an míbhuntáiste is mó a bhaineann le ASCII ná nach féidir leis ach litreacha an Bhéarla a léiriú. Dar ndóigh, ní raibh aon dul as ag lucht a dheartha ach géilleadh do smacht na seacht ngiotán mar bhí an teicneolaíocht teoranta ag an am. D'fhág sé sin oidhreacht throm ar thionscal na ríomhaireachta, áfach, oidhreacht nach bhfuil curtha dínn go hiomlán fiú sa lá atá inniu ann.

## *Breacadh an lae os cionn na hEorpa*

Tá na mílte teangacha ar domhan. Bíonn a hortagrafaíocht féin ag gabháil le gach teanga, agus ní fhreastalaíonn ASCII ach ar fhíorbheagán acu. Fiú gan dul thar theorainneacha na hEorpa, tá tuairim is 400 carachtar breisithe i ngnáthúsáid i dteangacha éagsúla sa mhullach ar an bhunaibítir Rómhánach. Séard is carachtar breisithe ann ná carachtar a bhfuil síneadh fada nó maisiúchán éigin eile ag gabháil leis: ä ö ü na Gearmáinise, á é í ó ú na Gaeilge, á é ě í ó ú ů č ď ň ř š ť ž na Seicise, agus mar sin de. Bíonn corrcharachtar go hiomlán nua le fáil in aibítreacha Rómhánacha chomh maith, mar shampla ß na Gearmáinise agus þ na hIoslannaise. Níl aon cheann díobh seo le fáil in ASCII, agus is mór an crá croí é. Cé gur minic gur fhorbair daoine bealach timpeall ar na srianta seo, mar shampla "ae oe ue" a scríobh in ionad "ä ö ü", nó "a/ o/ u/" a scríobh in ionad "á ó ú", ní réitigh shásúla iad go fadtéarmach.

Níos faide anonn faightear teangacha nach n-úsáideann an aibítir Rómhánach ar chor ar bith, dála na Rúisise agus na Bulgáirise, a bhfuil an aibítir Choireallach á húsáid acu, agus an Ghréigis,

teanga a bhfuil a haibítir féin aici. Feidhmíonn an aibítir Choireallach agus an aibítir Ghréagach ar aon dul leis an aibítir Rómhánach ach tá na carachtair difriúil. Tá difríochtaí níos bunúsaí i gceist leis an Araibis agus an Eabhrais: cé gur córais aibítreacha iad go bunúsach, is ó dheis go clé a scríobhtar iad, níl aon ghutaí iontu, agus níl aon difir idir ceannlitreacha agus litreacha beaga. Níl ASCII in ann téacs sna teangacha seo a thaifeadadh ar chor ar bith ach an aibítir a athscríobh go hiomlán san aibítir Rómhánach, rud nach mbíonn inghlactha do phobal labhartha na teanga de ghnáth, nó nach bhfuil indéanta fiú.

Agus ríomhairí ag éirí níos saoire an t-am ar fad, ní fada go raibh an scéal seo ina fhadhb agus gá ag daoine téacs ina dteanga féin a stóráil ar ríomhairí. Ar an dea-uair, bhí an teicneolaíocht in ann teacht i gcabhair. Faoi na naoi déag ochtóidí bhí gnáth-ailtireacht ríomhairí tar éis athrú agus ba é ocht ngiotán gnáthfhad an bhirt anois, rud a mhéadaigh faoi dhó ar an líon carachtar is féidir a thaifeadadh le tábla ionchódaithe amháin. Ba é 255 an uimhir a b'airde a d'fhéadfaí a stóráil in aon ghiotán amháin, rud a chuir ar chumas daoine táblaí ionchódaithe nua a chruthú agus 256 carachtar orthu seachas 128. Thug a lán comhlachtaí, dreamanna, daoine aonaracha fiú, faoi thograí chun a leagan féin de ASCII a chruthú – cuid acu ag freastal ar aon teanga amháin, cuid eile ag freastal ar ghrúpaí teangacha.

An iarracht ba rathúla ná sraith caighdeán a d'fhoilsigh an eagraíocht caighdeánaithe idirnáisiúnta ISO faoin teideal ISO 8859 sa bhliain 1987. Ní tábla ionchódaithe amháin atá ann ach cúpla ceann – glaoitear tacair carachtar orthu – agus gach ceann acu ag freastal ar ghrúpa teangacha. Mar shampla, freastalaíonn an tacar ISO 8859-1 ar theangacha iarthar na hEorpa, an tacar ISO 8859-2 ar theangacha lár agus oirthear na hEorpa a úsáideann an aibítir Rómhánach, ISO 8859-5 ar theangacha a úsáideann an aibítir Choireallach, agus mar sin de. Tá 15 tacar ann faoi láthair agus iad ag freastal go sásúil ar an chuid is mó de theangacha Eorpacha móide an Eabhrais, an Araibis agus an Téalainnis.

Tá ionchóduithe ISO 8859 leagtha amach le bheith comhoiriúnach le ASCII go siarghabhálach. Tá 256 carachtar i ngach tacar, ach tá an chéad 128 carachtar mar an chéanna le ASCII i gcónaí, fiú sna tacair Choireallacha, Arabacha, agus mar sin de. Is mar seo a dhéantar cinnte go bhféadfaí ar a laghad téacs ASCII a chur ó ríomhaire go ríomhaire gan truailliú, beag beann ar ionchódú.

Ó cruthaíodh caighdeán ISO 8859 agus caighdeáin eile mar é, réitíodh go leor fadhbanna ach, ag an am céanna, cruthaíodh fadhbanna nua. Tá an ríomhaireacht sna naoi déag ochtóidí agus naoi déag nóchaidí go mór faoi sciúirse an "téacs thruaillithe", mar a ghlaoitear air. Feiniméan é an téacs truaillithe a eascraíonn nuair a thaifeadann duine (nó ríomhaire) téacs de réir ionchódaithe áirithe agus nuair a osclaíonn duine nó ríomhaire eile faoi ionchódú difriúil é. Mar shampla, má

scríobhann an chéad duine an litir Sheiceach Š faoin ionchódú ISO 8859-2, taifeadtar é mar uimhir 169. Ansin, má osclaíonn an duine eile é faoin ionchódú ISO- 8859-5, gheobhaidh sé an litir Sheirbiach Љ, mar is é sin a litir a fhreagraíonn don uimhir 169 faoin ionchódú sin.

Bíonn cásanna den téacs truaillithe ag tarlú arís is arís eile ar fud an domhain, gach uair a osclaíonn duine ríomhphost, leathanach Gréasáin nó cáipéis faoin ionchódú mícheart. Mar gheall ar easpa eolais an úsáideora a tharlaíonn sé go minic, ach uaireanta eile is é an bogearra féin a bhíonn lochtach. Ní nach ionadh, mar tá an oiread sin scéimeanna ionchódaithe in úsáid anois gur deacair iad go léir a áireamh. Ní ISO amháin a thug faoi thacair carachtar a chruthú, chuir go leor dreamanna eile a ladar sa scéal chomh maith, idir chomhlachtaí tráchtála, eagrais oifigiúla agus eagraíochtaí deonacha. Is gá an comhlacht Microsoft a lua go háirithe, mar chinn an comhlacht seo gan cloí le caighdeáin ISO agus a gcuid n-ionchóduithe féin a chruthú – díreach mar a rinne IBM sna seascaidí agus iad ag tabhairt droim láimhe do ASCII. Is ó Microsoft a tháinig an téarma "códleathanach", téarma atá ar comhchiall le "tacar carachtar" a bheag nó a mhór. Tá cosúlachtaí móra idir caighdeáin Microsoft agus caighdeáin ISO: mar shampla, tá an códleathanach Windows-1252 an-chosúil leis an tacar carachtar ISO 8859-1, gan ach miondifríochtaí eatarthu.

Deacracht eile a bhaineann leis na tacair carachtar agus códleathanaigh ar fad seo ná go bhfuil sé deacair teangacha a mheascadh in aon phíosa téacs amháin. Abair go bhfuil gá agat le corr-abairt as Rúisis a lua istigh i dtéacs Gaeilge: níl aon ionchódú faoin spéir in ann an aibítir Choireallach agus gutaí fada na Gaeilge a láimhseáil ag an am céanna. Is gá ionchóduithe a mhalartú anonn is anall istigh sa téacs, rud atá casta agus mar thoradh air, níl mórán bogearraí in ann chuige.

Cé gur réiteach réasúnta sásúil é cur chuige na dtacar carachtar do theangacha iartharacha, ba bheag freastal a rinne sé ar theangacha na hÁise. Tá deacrachtaí faoi leith ag baint le leithéidí na Sínise agus na Seapáinise a chur ar ríomhaire. Ní teangacha aibítreacha iad na teangacha seo ar chor ar bith mar tá siombailí scríofa acu a sheasann d'fhocal nó do mhoirféim iomlán, agus tá na mílte de na siombailí seo ann. Ar feadh i bhfad, ba ghnách sna teangacha seo tacair carachtar a úsáid a bhí bunaithe ar dhá bheart (sé sin, 16 giotán) seachas ar bheart amháin (sé sin, ocht ngiotán), rud a mhéadaíonn ar an líon carachtar in-taifeadta suas go 65,536 (= $2^{16}$). Don tSínis, mar shampla, is líon sásúil é seo do na carachtair is coitianta, ach meastar go bhfuil timpeall is 80,000 carachtar sa teanga go hiomlán, cuid acu gan úsáid ach go hannamh.

## *Seo chugainn Unicode*

Ag féachaint siar ar na tacair, na códleathanaigh agus na deacrachtaí a bhaineann leo, is minic a ritheann an cheist seo le daoine: nach féidir tacar mór amháin a chruthú ina mbeadh na carachtair ar fad dá bhfuil ar domhan? An freagra ná gur féidir, agus rinneadh cheana é. An t-ainm atá ar an tacar carachtar sin ná Unicode. Is é atá in Unicode ná tábla ollmhór a bhfuil (nach mór) gach carachtar as (nach mór) gach teanga ar domhan liostaithe ann. Sé an *Unicode Consortium* an eagraíocht atá ag tabhairt cúraim do Unicode, eagraíocht neamhbhrabúis a bhfuil beagnach gach mórchomhlacht ríomhaireachta an domhain páirteach inti. Tá glactha ag ISO le Unicode mar chaighdeán dá gcuid féin chomh maith, faoin teideal ISO `10646` nó *Universal Character Set*.

Nuair a foilsíodh Unicode don chéad uair i `1991`, ba iad na spriocanna a bhí aige ná éalú ó na castachtaí go léir a bhaineann le hionchóduithe a mhalartú ó theanga go teanga. Amach anseo, ní bheadh ar dhaoine bheith buartha faoi ionchóduithe ar an ríomhphost agus ar shuímh Ghréasáin mar bheadh na teachtaireachtaí agus leathanaigh go léir san aon ionchódú amháin: Unicode.

Is buntáiste ollmhór do Unicode é a bheith ilteangach, ach ní hé sin an buntáiste is mó. An buntáiste is mó a bhaineann leis i gcomparáid leis na scéimeanna a tháinig roimhe ná nach bhfuil aon uasteorainn ag baint leis an líon carachtar. I bprionsabal, is féidir líon ar bith carachtar a liostáil ar Unicode, agus go deimhin, tá an *Unicode Consortium* ag glacadh le moltaí i gcónaí chun carachtair agus teangacha nua a chur ar an liosta, agus tá an tacar carachtar ag fás i gcónaí dá bharr. Conas a baineadh sin amach? Le bheith cruinn, ní ionchódú é Unicode, is tacar carachtar é: níl in Unicode ach tábla a shannaíonn uimhreacha do litreacha – glaoitear códphointe ar gach ceann de na huimhreacha a bhfuil litir sannaithe dó. Ní deir Unicode conas na huimhreacha a aistriú ina ngiotáin agus conas iad a stóráil i gcuimhne an ríomhaire. Fágtar an cúram sin faoi algartaim ionchódaithe. Ceann de na halgartaim ionchódaithe is coitianta ná `UTF-8`. Is algartam ilfhad é `UTF-8`: úsáideann sé beart amháin más charachtar Rómhánach ASCII atá á ionchódú, dhá bheart más carachtar Rómhánach breisithe é, agus suas go ceithre bheart do charachtair eile.

D'fhéadfaí a rá gur algartam comhbhrú é `UTF-8` (agus algartaim eile mar é) mar déanann sé cinnte nach nglacann aon charachtar níos mó spáis i gcuimhne an ríomhaire ná mar is gá. Mar shampla, ní theastaíonn ach beart amháin ó charachtair Rómhánacha ar nós `A` agus `M` mar tá códphointí faoi bhun `255` sannaithe dóibh in Unicode, agus dá bhrí sin, úsáideann `UTF-8` beart amháin orthu. Carachtar a bhfuil uimhir sannaithe dó a dteastaíonn breis is beart amháin uaidh, baineann `UTF-8` úsáid go díreach as an mhéid beart atá ag teastáil, gan dul thairis. Is féidir le `UTF-8` códphointí faoi bhun ceithre bheart a ionchódú, rud atá sásúil don mhéid carachtar atá

liostaithe ar Unicode faoi láthair. Má théann Unicode thar 4,294,967,296 (= $2^{4\times8}$) carachtar lá amháin amach anseo, beidh gá le UTF-8 a uasdátú, ach ní dócha go mbeidh gá leis sin go deo.

Tá glactha le Unicode go réasúnta forleathan i dtionscal na ríomhaireachta anois – oibríonn Windows ar Unicode go hinmheánach ó leagan 2000 i leith, mar shampla – cé nach beagbhríoch an jab é bogearra a chur in oiriúint dó. Ar an chéad dul síos, caithfidh an bogearra a bheith réidh le breis is beart amháin a úsáid chun carachtar a thaifeadadh, agus ar an dara dul síos, caithfidh eolas a bheith ag an bhogearra ar algartam éigin ionchódaithe ar nós UTF-8. Fós féin, is mór an faoiseamh é Unicode don ghnáthúsáideoir. Fad is a thacaíonn bogearra le Unicode, is féidir leis dul ag obair ar théacs i dteanga ar bith, gan an téacs a thruailliú ar an bhealach. An t-aon chúis go bhfuil an saol mór fós ag fulaingt mar gheall ar fheiniméan an téacs thruaillithe ná nach bhfuil glactha le Unicode forleathan go leor, agus go bhfuil fós go leor seancháipéisí agus seanbhogearraí amuigh ansin nach dtacaíonn leis, nó nach dtacaíonn leis go hiomlán.

## *Conclúid*

Cé nach bhfuil Unicode saor ó chonspóid, go háirithe ar cheisteanna a bhaineann le teangacha gan traidisiún láidir scríbhneoireachta, glactar leis anois gurb é Unicode an réiteach críochnúil ar na deacrachtaí a bhain le próiseáil téacs go dtí seo. Go deimhin, dá mbeadh Unicode againn ón chéad lá, ní bheadh aon deacracht le réiteach. Ach cé go raibh túsbhlianta na ríomhaireachta go mór faoi smacht ag saol an Bhéarla, is léir ón chuntas seo go ndeachaigh an tionscal i ngleic leis an fhadhb, agus go bhfuil ríomhairí á gcur in oiriúint, de réir a chéile, don iliomad teangacha atá ar domhan.

Anois agus tús na fiche is a haonú aoise buailte linn, níl fágtha ach a bheith ag súil go nglacfar le Unicode níos forleithne go mbeidh an bua aige ar scéimeanna easnamhacha ar nós ASCII agus ISO 8859. Ní fada uainn an tráth nuair a bheidh, i súile an ríomhaire ar a laghad, teangacha go léir an domhain ar comhchéim le chéile.