Το πρότυπο Unicode

Απόστολος Συρόπουλος

28ης Οκτωβρίου 366
671 00 Ξάνθη
E-mail: apostolo@obelix.ee.duth.gr

Δημήτριος Α. Φιλίππου

Κάτω Γατζέα
385 00 Βόλος

Ιωάννης Κ. Δημάκος
Παν/μιο Πατρών
E-mail: idimakos@upatras.gr

1. Εισαγωγή

Ένας απλός τρόπος για να κρίνει κανείς κατά πόσο είναι σοβαρό ή μη ένα σύστημα στοιχειοθεσίας είναι να ελέγξει τις δυνατότητες στοιχειοθεσίας μαθηματικού και πολυγλωσσικού κειμένου με τη χρήση του συστήματος αυτού. Μπορούμε εύκολα να καταλάβουμε γιατί το μαθηματικό κείμενο είναι ιδιαίτερα απαιτητικό αν σκεφτούμε μόνο ότι οι διάφοροι μαθηματικοί τύποι (εξισώσεις, κ.λπ.) είναι στην πραγματικότητα δισδιάστατο κείμενο. Από την άλλη, η σύνταξη ενός πολύγλωσσου κειμένου είναι δύσκολη γιατί πρέπει να υποστηρίζεται όχι μόνο ένας μεγάλος αριθμός γραμματοσειρών, αλλά και οι τυπογραφικές ιδιαιτερότητες της κάθε γλώσσας.

Τόσο το μαθηματικό κείμενο όσο και η υλοποίηση των τυπογραφικών ιδιαιτεροτήτων είναι θέματα στα οποία το TEX και οι επίγονοί του είναι πραγματικά πρωταθλητές. Όμως, το TEX είναι «γνήσιο Αμερικανάκι» και, σαν τέτοιο, βασίζεται στο αμερικανικό πρότυπο ASCII. Σύμφωνα με το πρότυπο ASCII, κάθε χαρακτήρας αντιστοιχεί σε έναν θετικό ακέραιο αριθμό του οποίου το μέγεθος δεν μπορεί να ξεπεράσει τα οκτώ bit (ένα byte).[1] Συνεπώς, κατά το πρότυπο ASCII, δεν μπορούμε να γράψουμε κάτι στον υπολογιστή σε μια γλώσσα που περιέχει περισσότερα από $2^8 = 256$ γράμματα και σύμβολα.

[1] Στην αρχική του μορφή, το ASCII περιελάμβανε μόνο «χαρακτήρες μεγέθους 7 bit», δηλαδή σε $2^7 = 128$ χαρακτήρες. Κατόπιν διάφοροι κατασκευαστές Η/Υ και λειτουργικών συστημάτων επέκτειναν το ASCII στα 8 bit (1 byte) — ο καθένας σύμφωνα με τον δικό του τρόπο.

Για το TEX, η λύση που επινοήθηκε για την αντιμετώπιση του προβλήματος της στοιχειοθεσίας κειμένων σε γλώσσες που έχουν περισσότερα από 256 γράμματα και σύμβολα ήταν η μεταγραφή, δηλαδή η γραφή ενός μη λατινικού αλφαβήτου χρησιμοποιώντας το λατινικό αλφάβητο στον κώδικα TEX. Το μειονέκτημα αυτής της μεθόδου είναι ότι το αρχείο με τον κώδικα TEX, γίνεται δυσανάγνωστο ακόμα και για τους μυημένους. Προβλήματα ακριβώς όπως αυτό, και κυρίως η ύπαρξη κωδικοποιήσεων (κωδικοσελίδων) για την ίδια γλώσσα διαφορετικών για κάθε λειτουργικό σύστημα, ήταν που οδήγησαν στη δημιουργία του προτύπου Unicode.

Με απλά λόγια, το πρότυπο Unicode είναι ένα μεγάλο σύνολο χαρακτήρων το οποίο υποστηρίζει τις περισσότερες απο τις γραφές του κόσμου (νεκρές ή και χρησιμοποιούμενες) επιτρέποντας έτσι την εύκολη μετάδοση, επεξεργασία και προβολή πολυγλωσσικών κειμένων. Σε ό,τι ακολουθεί θα παρουσιάσουμε σε συντομία τις τεχνικές λεπτομέρειες που αφορούν το πρότυπο, τις ομάδες χαρακτήρων που υποστηρίζει αλλά και την ελληνική υποστήριξη.

2. Περιγραφή του προτύπου Unicode

Το πρότυπο Unicode δημιουργήθηκε με βάση μια σειρά από αρχές. Κύρια αρχή αποτέλεσε η απαίτηση ο κάθε χαρακτήρας να αναπαριστάται από 16 bit (δηλαδή δύο byte). Ο κάθε χαρακήρας αντιστοιχεί σε ένα *κωδικό σημείο* το οποίο αναπαριστά τη θέση του χαρακτήρα στο πρότυπο και μπορεί να χρησιμοποιηθεί για συγκρίσεις. Το κωδικό σημείο κάθε χαρακτήρα σημειώνεται με ένα τετραψήφιο δεκαεξαδικό αριθμό, π.χ., AF01, 0308, κ.ο.κ. Επιπλέον, τα αρχεία Unicode θα έπρεπε να είναι εύκολα προσπελάσιμα, ενώ η επεξεργασία τους δεν θα έπρεπε να παρουσιάζει ιδιαίτερες δυσκολίες. Επειδή το Unicode είναι ουσιαστικά ένα (μεγάλο) σύνολο χαρακτήρων, ακριβώς όπως είναι και το ASCII, ο αναγνώστης θα πρέπει να έχει στο μυαλό του ότι το πρότυπο Unicode καθορίζει τον αριθμό στον οποίο αντιστοιχεί ένας χαρακτήρας η ένας *γλύφος* (δηλαδή ένας χαρακτήρας που προκύπτει από τον συνδυασμό δύο ή περισσοτέρων απλών χαρακτήρων), αλλά δεν καθορίζει διαφορετικές μορφές (σχεδιάσεις) χαρακτήρων. Για παράδειγμα, στο παρακάτω σχήμα δίνουμε τέσσερις διαφορετικές μορφές (σχεδιάσεις) του γλύφου LATIN SMALL LIGATURE FI, που αντιστοιχεί στον κωδικό αριθμό Unicode FB01:

ﬁ ﬁ ﬁ ﬁ

Επιπλέον, προς αποφυγή προβλημάτων ανάμειξης διαφορετικών γλωσσών, αρκετοί χαρακτήρες που μοιάζουν εμφανίζονται πολλές φορές. Μια τέτοια περίπτωση είναι ο χαρακτήρας A που απαντάται τουλάχιστον στις εξής μορφές: LATIN CAP-

ITAL LETTER A, GREEK CAPITAL LETTER ALPHA και CYRILLIC CAPITAL LETTER A.

Το πρότυπο Unicode έχει σχεδιασθεί έτσι ώστε ο κάθε χαρακτήρας να αντιστοιχεί σε μία σειρά 16 bit. Αυτή η σύμβαση οδηγεί στο συμπέρασμα ότι με το Unicode μπορούμε να αναπαραστήσουμε μέχρι $2^{16} = 65.536$ διαφορετικούς χαρακτήρες. Όμως στην πραγματικότητα μπορούμε άμεσα να απαραστήσουμε 63.486 χαρακτήρες, ενώ οι υπόλοιπες 2.048 δεκαεξάδες χρησιμοποιούνται για την αναπαράσταση άλλων 1.048.544 χαρακτήρων χρησιμοποιώντας ζεύγη δεκαεξάδων τα οποία ονομάζονται *αντικαταστάτες* (surrogates). Έτσι, το Unicode μάς επιτρέπει να αναπαραστήσουμε 1.112.030 διαφορετικούς χαρακτήρες.

Για την αναπαράσταση ενός χαρακτήρα του Unicode, χρησιμοποιούμε όλα τα 256 διαφορετικά byte. Αυτό όμως έχει το μειονέκτημα ότι αν έχουμε μια εφαρμογή η οποία αναμένει κάποια συγκεκριμένα byte και αυτή πρέπει να επεξεργαστεί ένα αρχείο Unicode, τότε είναι πολύ πιθανό να αποτύχει. Έτσι για λόγους συμβατότητας το πρότυπο Unicode ορίζει και μια μορφή κωδικοποίησης στην οποία χρησιμοποιούμε οκτάδες bit αντί για δεκαεξάδες για την αναπαράσταση των χαρακτήρων. Η μορφή αυτή είναι γνωστή ως UTF-8, ενώ η εξ ορισμού μορφή των 16 bit είναι γνωστή ως UTF-16.

Στην κωδικοποίηση UTF-8 οι λατινικοί χαρακτήρες που περιέχονται στο πρότυπο ASCII αναπαριστώνται όπως πριν, δηλαδή με ένα byte. Για την αναπαράσταση των ιαπωνικών, κινεζικών και κορεατικών χαρακτήρων χρησιμοποιούνται τρία byte, ενώ για τους χαρακτήρες που απαντώνται στα απλά φθογγικά αλφάβητα (π.χ., το ελληνικό, το κυριλλικό, κ.ά.) χρησιμοποιούνται δύο byte. Τέλος, για την αναπαράσταση των επιπλέον χαρακτήρων χρησιμοποιούνται τέσσερα byte.

Η κωδικοποίηση UTF-16 όταν δεν χρησιμοποιούνται οι αντικαταστάτες, είναι γνωστή ως κωδικοποίηση UCS-2. Ακομή, επειδή υπάρχουν δύο τρόποι για την αποθήκευση των bit σ' ένα byte, θα πρέπει είτε ο πρώτος χαρακτήρας του αρχείου μας να είναι ο ZERO WIDTH NO-BREAK SPACE, είτε να καθορίζουμε εμείς τον τρόπο κωδικοποίησης των byte. Στην περίπτωση που που τα λιγότερο σημαντικά bit βρίσκονται στο τέλος, λέμε ότι τα byte έχουν μορφή big endian, αλλιώς ότι έχουν μορφή little endian. Για να κατάλαβεται τη διαφορά, σας λέμε ότι αν τα byte έχουν μορφή big endian το γράμμα C αναπαριστάνεται με τα ψηφία 00000000 01000011, ενώ στην άλλη περίπτωση με τα ψηφία 00000000 01100001. Έτσι έχουμε τις κωδικοποίησεις: UTF-16BE, UTF-16LE, UCS-2BE και UCS-2LE.

Οι αντικαταστάτες σχηματίζονται από δύο ζεύγη byte τα οποία ονομάζονται *άνω αντικαταστάτης* (high surrogate) και *κάτω αντικαταστάτης* (low surrogates). Ο άνω αντικαστάτης θα πρέπει να αντιστοιχεί σ' ένα κωδικό σημείο μεταξύ του D800 και του DBFF, ενώ ο κάτω αντικαταστάτης θα πρέπει να αντιστοιχεί σ' ένα κωδικό σημείο μεταξύ DC00 και DFFF.

3. Οι χαρακτήρες που «υποστηρίζει» το Unicode

Σ' αυτή την ενότητα παρουσιάζουμε συνοπτικά τους διάφορους χαρακτήρες που υποστηρίζει προς το παρόν το Unicode. Κατ' αρχάς, οι γραφές που υποστηρίζονται είναι: αραβική, αρμενική, Bengali, Bopomofo, Buhid, καναδική συλλαβική γραφή, τσερόκι, κυριλλική, Deseret, Devanagari, αιθιοπική, γεωργιανή, γοτθική, ελληνική, Gujarati, Gurmukhi, κινεζική (ή Χαν, όπως είναι το σωστό), κορεατική, εβραϊκή, ιαπωνική (συλλαβική και ιδεογραφική), χμέρ, λατινικές γραφές, γραφή της γλώσσας του Λάος, Malayalam, μογγολικά, Myanmar, Ogham, ετρουσκική, Oriya, ρουνική, γι, θιβετιανή, ταϋλανδέζική, κ.ά.[2]

Επιπλέον, υπάρχουν και αρκετά σύμβολα τα οποία δεν είναι γράμματα και περιλαμβάνονται στο πρότυπο Unicode. Τα σύμβολα αυτά είναι: αριθμοί, τονικά σύμβολα, σύμβολα στίξης, γενικά σύμβολα, μαθηματικά σύμβολα, μουσικά σύμβολα (δυτικά και βυζαντινά), τεχνικά σύμβολα, βέλη, πλαίσια και άλλα γεωμετρικά σύμβολα, τα σύμβολα του αλφαβήτου Braille για τους τυφλούς και τα Kangxi radicals που χρησιμοποιούνται στα κινεζικά λεξικά μιας και δεν είναι δυνατό να υπάρχουν χιλιάδες ενότητες.

Αν και το Unicode υποστηρίζει όλες τις γραφές που παρουσιάσαμε παραπάνω, είναι επιπλέον συμβατό και με τα διάφορα πρότυπα 8-bit, όπως για παράδειγμα το Latin-1, το ISO8859-7, κ.λπ. Ο Πίνακας 1 δείχνει την κωδικοποίηση της ελληνικής μονοτονικής γραφής: Οι χαρακτήρες που δεν φαίνονται είναι απλά εναλλακτικές μορφές διαφόρων γραμμάτων (π.χ., π και ϖ) οι οποίες όμως δεν υποστήριζονται από τη γραμματοσειρά που χρησιμοποιούμε.

Στον Πίνακα 2 στο τέλος του άρθρου, παρουσιάζουμε την κωδικοποίηση της ελληνικής πολυτονικής γραφής κατά το πρότυπο Unicode.

Αξίζει να αναφέρουμε ότι οι πίνακες ετούτου του άρθρου στοιχειοθετήθηκαν με το Λ και μια εν δυνάμει γραμματοσειρά τύπου Unicode η οποία στηρίζεται στις γραμματοσειρές που χρησιμοποιούνται στο πακέτο babel. Η γραμματοσειρά είναι τμήμα μιας σειράς εργαλείων τα οποία ελπίζουμε να διαθέσουμε προσεχώς στο ευρύ κοινό.

Περισσότερες πληροφορίες σχετικά με το πρότυπο Unicode μπορείτε να βρείτε στο δικτυακό τόπο: `http://www.unicode.org`.

Η Συντακτική Επιτροπή του περιοδικού θα ήθελε να προσκαλέσει όλους τους αναγνώστες του περιοδικού (αλλά και άλλους ενδιαφερόμενους) να μελετήσουν τους πίνακες του πρότυπου Unicode και να παρουσιάσουν τις δικές τους απόψεις. Οι συγγραφείς εύχονται το άρθρο τους και άλλα παρόμοια που έχουν δημοσιευτεί στο *Εὔτυπον* να δώσουν την αφορμή για μια γόνιμη και δημιουργική ανταλλαγή

[2] Οι γραφές που αναφέρουμε εδώ με λατινικά γράμματα είναι γραφές που δεν τις έχουμε συναντήσει ποτέ στα ελληνικά. Απαντώνται κυρίως στην Ινδική Χερσόνησο.

0374	ʹ	0375	͵	037A	ͺ	037E	;	0386	Ά
0388	Έ	0389	Ή	038A	Ί	038C	Ό	038E	Ύ
038F	Ώ	0390	ΐ	0391	Α	0392	Β	0393	Γ
0394	Δ	0395	Ε	0396	Ζ	0397	Η	0398	Θ
0399	Ι	039A	Κ	039B	Λ	039C	Μ	039D	Ν
039E	Ξ	039F	Ο	03A0	Π	03A1	Ρ	03A3	Σ
03A4	Τ	03A5	Υ	03A6	Φ	03A7	Χ	03A8	Ψ
03A9	Ω	03AA	Ϊ	03AB	Ϋ	03AC	ά	03AD	έ
03AE	ή	03AF	ί	03B0	ΰ	03B1	α	03B2	β
03B3	γ	03B4	δ	03B5	ε	03B6	ζ	03B7	η
03B8	ϑ	03B9	ι	03BA	κ	03BB	λ	03BC	μ
03BD	ν	03BE	ξ	03BF	ο	03C0	π	03C1	ρ
03C2	ς	03C3	σ	03C4	τ	03C5	υ	03C6	φ
03C7	χ	03C8	ψ	03C9	ω	03CA	ϊ	03CB	ϋ
03CC	ό	03CD	ύ	03CE	ώ	03D0		03D1	
03D2		03D3		03D4		03D5		03D6	
03D7	ϗ	03DA	Ϛ	03DB	ϛ	03DC	F	03DD	ϝ
03DE	Ϟ	03DF	ϟ	03E0	Ϡ	03E1	ϡ		

Πίνακας 1: Η κωδικοποίηση των χαρακτήρων της ελληνικής μονοτονικής γραφής κατά το πρότυπο Unicode.

1FB8	Ᾰ	1FB9	Ᾱ	1FBA	Ὰ	1FBB	Ά	1FBC	Α$_{\text{ι}}$
1FBD		1FC2	ῂ	1FC3	ῃ	1FC4	ή	1FC6	ῆ
1FC7	ῇ	1FC8	Ὲ	1FC9	Έ	1FCA	Ὴ	1FCB	Ή
1FCC	Η$_{\text{ι}}$	1FD0	ῐ	1FD1	ῑ	1FD2	ῒ	1FD3	ΐ
1FD6	ῖ	1FD7	ῗ	1FD8	Ῐ	1FD9	Ῑ	1FDA	Ὶ
1FDB	Ί	1FE0	ῠ	1FE1	ῡ	1FE2	ῢ	1FE3	ΰ
1FE4	ῤ	1FE5	ῥ	1FE6	ῦ	1FE7	ῧ	1FE8	Ῠ
1FE9	Ῡ	1FEA	Ὺ	1FEB	Ύ	1FEC	Ῥ	1FF2	ῲ
1FF3	ῳ	1FF4	ῴ	1FF6	ῶ	1FF7	ῷ	1FF8	Ὸ
1FF9	Ό	1FFA	Ὼ	1FFB	Ώ	1FFC	Ω$_{\text{ι}}$		

Πίνακας 2: Η κωδικοποίηση των χαρακτήρων της ελληνικής πολυτονικής γραφής κατά το πρότυπο Unicode.

απόψεων και προτάσεων για την περαιτέρω ορθότερη ανάπτυξη του πρότυπου αυτού όσον αφορά την ελληνική γλώσσα.

1F00	ἀ	1F01	ἁ	1F02	ἂ	1F03	ἃ	1F04	ἄ
1F05	ἅ	1F06	ἆ	1F07	ἇ	1F08	Ἀ	1F09	Ἁ
1F0A	Ἂ	1F0B	Ἃ	1F0C	Ἄ	1F0D	Ἅ	1F0E	Ἆ
1F0F	Ἇ	1F10	ἐ	1F11	ἑ	1F12	ἒ	1F13	ἓ
1F14	ἔ	1F15	ἕ	1F18	Ἐ	1F19	Ἑ	1F1A	Ἒ
1F1B	Ἓ	1F1C	Ἔ	1F1D	Ἕ	1F20	ἠ	1F21	ἡ
1F22	ἢ	1F23	ἣ	1F24	ἤ	1F25	ἥ	1F26	ἦ
1F27	ἧ	1F28	Ἠ	1F29	Ἡ	1F2A	Ἢ	1F2B	Ἣ
1F2C	Ἤ	1F2D	Ἥ	1F2E	Ἦ	1F2F	Ἧ	1F30	ἰ
1F31	ἱ	1F32	ἲ	1F33	ἳ	1F34	ἴ	1F35	ἵ
1F36	ἶ	1F37	ἷ	1F38	Ἰ	1F39	Ἱ	1F3A	Ἲ
1F3B	Ἳ	1F3C	Ἴ	1F3D	Ἵ	1F3E	Ἶ	1F3F	Ἷ
1F40	ὀ	1F41	ὁ	1F42	ὂ	1F43	ὃ	1F44	ὄ
1F45	ὅ	1F48	Ὀ	1F49	Ὁ	1F4A	Ὂ	1F4B	Ὃ
1F4C	Ὄ	1F4D	Ὅ	1F50	ὐ	1F51	ὑ	1F52	ὒ
1F53	ὓ	1F54	ὔ	1F55	ὕ	1F56	ὖ	1F57	ὗ
1F59	Ὑ	1F5B	Ὓ	1F5D	Ὕ	1F5F	Ὗ	1F60	ὠ
1F61	ὡ	1F62	ὢ	1F63	ὣ	1F64	ὤ	1F65	ὥ
1F66	ὦ	1F67	ὧ	1F68	Ὠ	1F69	Ὡ	1F6A	Ὢ
1F6B	Ὣ	1F6C	Ὤ	1F6D	Ὥ	1F6E	Ὦ	1F6F	Ὧ
1F70	ὰ	1F71	ά	1F72	ὲ	1F73	έ	1F74	ὴ
1F75	ή	1F76	ὶ	1F77	ί	1F78	ὸ	1F79	ό
1F7A	ὺ	1F7B	ύ	1F7C	ὼ	1F7D	ώ	1F80	ᾀ
1F81	ᾁ	1F82	ᾂ	1F83	ᾃ	1F84	ᾄ	1F85	ᾅ
1F86	ᾆ	1F87	ᾇ	1F88	ᾈ	1F89	ᾉ	1F8A	ᾊ
1F8B	ᾋ	1F8C	ᾌ	1F8D	ᾍ	1F8E	ᾎ	1F8F	ᾏ
1F90	ᾐ	1F91	ᾑ	1F92	ᾒ	1F93	ᾓ	1F94	ᾔ
1F95	ᾕ	1F96	ᾖ	1F97	ᾗ	1F98	ᾘ	1F99	ᾙ
1F9A	ᾚ	1F9B	ᾛ	1F9C	ᾜ	1F9D	ᾝ	1F9E	ᾞ
1F9F	ᾟ	1FA0	ᾠ	1FA1	ᾡ	1FA2	ᾢ	1FA3	ᾣ
1FA4	ᾤ	1FA5	ᾥ	1FA6	ᾦ	1FA7	ᾧ	1FA8	ᾨ
1FA9	ᾩ	1FAA	ᾪ	1FAB	ᾫ	1FAC	ᾬ	1FAD	ᾭ
1FAE	ᾮ	1FAF	ᾯ	1FB0	ᾰ	1FB1	ᾱ	1FB2	ᾲ
1FB3	ᾳ	1FB4	ᾴ	1FB6	ᾶ	1FB7	ᾷ		

Πίναχας 2: συνέχεια.