Met XeTeX meertalig
Talen en fonts in TeX

Abstract
Dit artikel is een bewerking van de lezing die ik op de NTG voorjaarsbijeenkomst van 8 juni 2007 heb gehouden. Dit artikel begint met wat achtergrondinformatie over talen, schriften en fonts. Het tweede deel van het artikel geeft een aantal voorbeelden van meertalig TeX-gebruik met behulp van XeTeX.

Keywords
talen, schrift, fonts, XeTeX

Talen

Als we TeX gebruiken doen we dat voor het overbrengen van informatie die in een bepaalde code weergegeven is. Deze code (meestal noemen we die code 'taal') omvat zowel de natuurlijke taal die we als mensen in het sociale verkeer gebruiken, als de kunstmatige, specialistische talen die bijvoorbeeld wiskundigen gebruiken om een n-dimensionale ruimte mee te beschrijven. In dit artikel beperk ik mij tot natuurlijke, levende talen, de talen die door mensen op dit moment gesproken worden. Hoeveel talen zijn er? Ik ga gemakshalve voorbij aan de lastige definitiekwestie van taal versus dialect en aan taalvariatie die het tellen van talen soms lastig maakt. Volgens de Ethnologue[1] worden er op dit moment ongeveer 6912 talen gesproken. Het woord 'ongeveer' staat voor: het aantal is niet precies want er worden nog steeds nieuwe talen ontdekt, terwijl er ook talen uitsterven. Dat er toch een getal staat komt omdat de gegevens uit een database komen waarin nu eenmaal 6912 records zitten van evenzoveel talen waarvan in elk geval een minimum aantal gegevens beschikbaar zijn. Een van de dingen die het meeste opvalt als het gaat om de aantallen is de enorme onbalans tussen de hoeveelheid talen en de hoeveelheid sprekers van die talen. Er zijn 10 talen met meer dan 100 miljoen sprekers: Chinees, Spaans, Engels, Arabisch, Hindi, Portugees, Bengaals, Russisch, Japans en Duits. In totaal zijn er 347 talen (ongeveer 5% van het totale aantal talen) met meer dan één miljoen sprekers. Deze talen worden gesproken door 94% van de wereldbevolking. De overige 95% van de talen worden gesproken door slechts 6% van de wereldbevolking. Zo'n 3900 van die talen heeft minder dan 10.000 spekers. Als de sprekers van die talen hun taal niet kunnen gaan schrijven is de kans groot dat zulke talen uitsterven en de cultuur, waarvan die taal een integraal onderdeel is, verloren gaat. Er zijn verschillende redenen waarom talen (nog) niet geschreven worden. Een belangrijke oorzaak is de orale cultuur waarin informatieoverdracht door middel van vertelde verhalen de norm is en schriftelijke communicatie de uitzondering, of zelfs gewoon niet bestaat. Een van de belangrijkste technische obstakels is de digitale kloof die het schrijven met behulp van een computer voor veel mensen onmogelijk maakt. Dat kan komen doordat er geen (vrije) fonts beschikbaar zijn of doordat het gebruikte schrift dermate complex is dat een 'standaard-oplossing' geen oplossing is. Daarnaast is de beperkte beschikbaarheid of de afwezigheid van lees- en schrijfonderwijs in de moedertaal voor sprekers van veel minderheidstalen een extra barrière.

Taal, schrift en font

Een paar keer is het al gegaan over 'taal', 'schrift' en 'font', maar wat is wat precies en wat is het verschil tussen het een en het ander? Een taal is niet hetzelfde als een schrift en een schrift is niet hetzelfde als een font. Toen mensen eenmaal hun taal gingen schrijven gebruikten ze karakters om de spraakklanken op papier vast te leggen. Deze karakters vormen samen met bijvoorbeeld de leestekens een 'schrift' of schriftsoort. Wij gebruiken voor de Nederlandse taal het Latijnse schrift, maar we zouden het ook best in het Cyrillisch of het Arabisch(-e schrift) kunnen schrijven. Op dezelfde manier kan een Arabische (taal) tekst in het Latijnse schrift geschreven worden. De combinatie van een schrift en een taal heet in het jargon een schriftsysteem (writing system). Dit is een combinatie van alle karakters, samen met de regels die beschrijven hoe die karakters gebruikt moeten worden in die specifieke taal. De minderheidstalen van de wereld maken vaak gebruik van het schrift van de meerderheidstaal in hun omgeving. Omdat er in zo'n minderheidstaal bijvoorbeeld andere klanken voorkomen, is het soms nodig om symbolen of accenttekens aan het schrift toe te voegen die in de meerderheidstaal niet nodig zijn. Ook kan het zijn dat bepaalde karakters of tekens op een andere manier gebruikt worden dan in de meerderheidstaal. Om die reden is er vaak per taal een verschillend 'schriftsysteem' te beschrijven, zelfs al zijn de (meeste) tekens hetzelfde. Voor zover we weten zijn er minstens zo'n 150 verschillende schriften. Zestig daarvan worden niet meer gebruikt (denk daarbij aan het spijkerschrift of aan Egyptische hiëroglyfen.) Negentig schriften worden op dit moment wel gebruikt. Ongeveer 50 daarvan zijn opgenomen in Unicode, de internationale standaard voor schrift-gebruik, 40 (nog) niet. Een font is een stukje software waarin onder andere de vorm van de verschillende karakters staat beschreven. Verder kunnen fonts informatie bevatten over hoe een karakter in een bepaald schriftsysteem gebruikt moet worden. Met name voor schriften als het Arabisch en de Aziatische schriften zijn deze 'rendering smarts' onmisbaar voor de juiste weergave van de gebruikte karakters. Een font kan dus verschillende schriftsystemen ondersteunen. Tenslotte is er nog het onderscheid tussen een karakter en een glyph. Het begrip karakter staat voor een abstract betekenisdragend element, zoals dat bijvoorbeeld in de Unicode tabel is beschreven. Het karakter 'a' is 'Latijn kleine letter a' in Unicode-termen. Een glyph is de specifieke vorm die de kleine letter 'a' krijgt in een bepaald font. Op een abstract niveau is het hetzelfde (een karakter 'a') maar op het niveau van de weergave ziet het er verschillend uit. Vergelijk bijvoorbeeld de 'a' in het woord meertalig in de titel boven dit artikel met de 'a' in het woord 'talen' in de ondertitel. Dit zijn verschillende glyphs voor hetzelfde karakter.

Schrijven wereldwijd

In de loop van de tijd zijn steeds meer mensen hun taal gaan schrijven. Er zijn verschillende routes om het fonetische materiaal (de spraakklanken) om te zetten in codes op papier (of in klei, of op papyrus of op boomschors). Ons eigen Romeinse (of Latijnse) schrift is een voorbeeld van een alfabetisch schrift waarbij iedere klank ongeveer een karakter heeft. Met behulp van een paar extra tweeklanken (eu, ui) of accenttekens kunnen we de gesproken taal redelijk goed weergeven. Nu zit er in de manier waarop wij onze taal schrijven een behoorlijke redundantie. "Drm s ht gd mglk m d klnkrs wg t ltn. Zekr ls je mt wt leeshulpn knt werkn". Arabisch en Hebreeuws zijn voorbeelden van zo'n manier van schrijven. Het gebruik van klinkertekens als een soort leeshulp om de vastgestelde tekst duidelijk te maken laat overigens wel zien dat het toch ook weer niet zo gek is om de klinkers toch op te schrijven.

Voorbeeld 1: arabisch schrift

r.1 العربي

r.2 يبرعلا

Het Arabisch is een schrift waarin de karakters afhankelijk van de plaats in het woord een verschillende vorm hebben. Daarnaast smelten opeenvolgende karakters samen tot zogenaamde ligaturen. Het bovenstaande voorbeeld laat dat zien: regel 1 bevat de afzonderlijke karakters, regel 2 de 'aan elkaar geschreven' vorm.

Overigens is het Arabische schrift ook een voorbeeld van het feit dat schrift niet alleen te maken heeft met techniek, maar ook met cultuur en geschiedenis. Het Arabische schrift kent een lange, rijke traditie van vormen en variatie in de kalligrafie (het met de hand geschreven schrift.) Een van de debatten in de wereld van Arabische fontontwerpers gaat over de vraag of er voor het Arabisch een zelfstandige typografie (door een machine geschreven schrift) mogelijk en wenselijk is. Sommigen menen dat fonts de traditionele kalligrafische vormen en variaties tot in de details moeten volgen. Volgens anderen is er ruimte voor een duidelijke eigen stijl in de fonts die nu ontworpen worden.

Voorbeeld 2 : aziatische schriften

Helemaal aan het andere uiterste van het spectrum zitten benaderingen van schrift die niet uitgaan van de vorm (de klanken) maar van de betekenis (de abstracte concepten), zoals bijvoorbeeld pictogrammen. Schriftsoorten zoals het Japans of Chinees vallen in deze categorie. Weer een andere benadering komt voor in verschillende Zuid-Oost Aziatische talen, zoals bijvoorbeeld het Khmer. Hier is de basis voor het schriftsysteem een medeklinker-klinker-groep. Dit komt ongeveer overeen met de lettergrepen. Er is dus een karakter voor de 'da', voor 'do', voor 'ro', etcetera. Als je met zo'n schrift de klankencombinatie 'door' wilt weergeven schrijf je dus: 'doro'. Voor ons gevoel lijkt het alsof de letters o en r van plaats zijn veranderd, in werkelijkheid zijn ze voor dit schrift op de goede plek geschreven, iedere Khmer-lezer begrijpt dat hier 'door' staat. Hieronder staat nog een voorbeeld van deze 'reordering', nu in het Devanagari schrift.

ब + ि + र + म = बिर्म

b + i + r + ma = birma

Model voor complex schrift

Als we kijken hoe ons eigen Romeinse schrift in de computer verwerkt wordt, dan is er sprake van een redelijk overzichtelijk, eenvoudig model: er is een één-op-één relatie tussen de toetsaanslag en het geheugen en tussen het geheugen en de uitvoer (beeldscherm/printer.) Om de diversiteit aan complexe, niet-Romeinse schriften goed te kunnen verwerken is een ingewikkelder model nodig. Aan de invoerkant moet er vaak gewerkt worden met Input Method Editors. Dit zijn hulpmiddelen die het mogelijk maken om karakters uit een schrift met duizenden karakters te kiezen. Dit kan bijvoorbeeld door middel van een apart invoerscherm of door speciale software die een fonetische transcriptie omzet in de gewenste karakters. In het geheugen worden de gegevens opgeslagen als een verzameling Unicode-karakters. Ook aan de uitvoerkant moet een computer om weten te gaan met speciale eisen die een complex schrift stelt, zoals bijvoorbeeld de volgordeverandering in het Devanagari-voorbeeld hierboven.

Enter XeTeX

XeTeX is een uitbreiding aan de standaard TeX engine die alle systeemfonts eenvoudig toegankelijk maakt voor de gebruikers. XeTeX werkt op Mac OS X, Linux en Windows en zit standaard in TeXLive 2007. Voor de meeste gebruikers is de belangrijkste reden om XeTeX te gebruiken de toegankelijkheid van de systeemfonts. Alle OT, TT en PS fonts die in een willekeurig tekstverwerkingsprogramma te gebruiken zijn, werken ook met XeTeX. Daarbij kunnen de font-setup troubles die zo kenmerkend zijn voor TeX achterwege blijven. Ook kunnen slimme fonts gebruikt worden in combinatie met speciale layout-engines die er voor zorgen dat de gebruikte fonts zich gedragen volgens de regels die gelden voor het betreffende schriftsysteem. Voorbeelden hiervan zijn AAT (Apple's Advanced Typography)[2] op de Mac, of het gebruik van de ICU-library[3] en SIL's Graphite techniek[4]. XeTeX is in de TeXwereld enthousiast ontvangen. In LaTeX zijn er verschillende packages gemaakt die XeTeX ondersteunen, waarvan het fontspec-package wel de belangrijkste is. Verder herkennen veel gebruikte packages zoals graphics, hyperref en PStricks het gebruik van XeTeX automatisch. Ook ConTeXt kan prima met XeTeX overweg[5]. Waar luaTeX met behulp van de ingebouwde scriptingtaal Lua de TeX-engine toegankelijk maakt voor bijvoorbeeld OT support, gaat XeTeX een andere weg: het maakt gebruik van libraries op besturingssysteemniveau zoals fontconfig (voor het vinden van de fonts) en de al genoemde AAT, ICU en Graphite.

Voorbeeld 3 : Typografische features in OT fonts

Hier volgen een paar voorbeelden van hoe XeTeX typografische features van OpenType fonts toegankelijk maakt. De tekst "Hallo Wereld! 0123456789" kunnen we in het font Garamond Premiere Pro zetten (voorbeeld 1). Met de aanduiding +smcp worden de letters als klein kapitalen gezet (voorbeeld 2) en met de code +sups selecteren we de optie superscript, wat voor dit font blijkbaar niet werkt voor hoofdletters en leestekens.

standaard:"GaramondPremrPro" Hallo Wereld! 0123456789
var1: "GaramondPremrPro:+smcp" HALLO WERELD! 0123456789
var2: "GaramondPremrPro:+sups" H^allo W^ereld! ^0123456789

Een ander voorbeeld van stylistische variaties, deze keer met Apple's AAT techniek, laat het volgende set met variaties van het Apple Chancery font zien.

"Apple Chancery:Design Complexity=Simple Design Level" at 16pt.
"Apple Chancery:Letter Case=Small Caps" at 16pt.
"Apple Chancery:Design Complexity=Flourishes Set A" at 16pt.
"Apple Chancery:Design Complexity=Flourishes Set B" at 16pt.
"Apple Chancery:Design Complexity=Flourishes Set C" at 16pt.

Voorbeeld 4 : Taal en schrift features in XƎTEX

Met de Unicode ondersteuning kan XƎTEX het ideaal van ‘iedere taal en ieder schrift’ ook voor TEX-gebruikers een stapje dichterbij brengen. Hier volgen een paar voorbeelden van taal- en schrift-specifieke features.

Speciale karakters zoals de Bengaalse klinker O worden geschreven met twee los van elkaar staande tekens die samen de medeklinker insluiten. Het voorbeeld hieronder geeft eerst de afzonderlijke karakters zonder speciale weergave, daarna de juiste, samengevoegde weergave die te activeren is met de code: ’script=beng’.

zonder 'script' aanduiding

<LETTER KA> : ক <VOWEL O> : ো → কো

met 'script=beng'

<LETTER KA> : ক <VOWEL O> : ো → কো

Het font Charis SIL[6] is een voorbeeld van een font met speciale taalspecifieke coderingen. In het onderstaande voorbeeld van een tekst in het Vietnamees is het verschil in de plaatsing van de accenttekens te zien. Deze speciale plaatsing is van belang voor de juiste weergave van het Vietnamees.

Unicode cung cấp một con số duy nhất cho mỗi
Unicode cung cấp một con số duy nhất cho mỗi

In het kader van dit korte artikel is het onmogelijk om een overzicht te geven van alle verschillende soorten schriften en manieren om daar TEXnisch mee om te gaan. Hopelijk maken deze paar voorbeelden in elk geval duidelijk dat er op dit gebied veel mogelijk is.

Noten

1. Raymond G. Gordon, ed., Ethnologue: Languages of the World, 15th edition, 2005. Zie voor de digitale versie: www.ethnologue.com
2. Zie ook: developer.apple.com/textfonts/
3. Zie ook: www.icu-project.org/
4. Zie ook: scripts.sil.org/RenderingGraphite
5. In elk geval meestal, zie ook: wiki.contextgarden.net/XeTeX
6. Zie ook: scripts.sil.org/CharisSILfont

Jelle Huisman
SIL International
jelle_huisman (at) sil (dot) org