

# Koding av samiske tekster

## Table of contents

|                                      |   |
|--------------------------------------|---|
| 1 De kjente problemene.....          | 2 |
| 2 Løsningen.....                     | 2 |
| 3 Denne portalen bruker Unicode..... | 2 |
| 4 Referanser.....                    | 3 |

## 1. De kjente problemene

Alle kjenner problemene med samisk og data: det er ikke mulig å skrive inn samiske tegn, og om man får en samisk tekst, er det nesten garantert at ett eller flere samiske tegn enten er borte eller kommer frem som noe helt annet. Det underliggende tekniske problemet er at det finnes flere inkompatible standarder for hvordan samiske tegn skal kodes, dvs hvilket nummer hvert tegn skal ha (en datamaskin skjønner egentlig bare nummer). I tillegg har de løsningene som har eksistert delvis vært private, eller i det minste ikke inkludert i operativsystemene til de store leverandørene, slik at det i praksis har vært svært vanskelig eller dyrt (eller begge deler) for vanlige brukere å få samisk til å fungere på datamaskinen.

## 2. Løsningen

Løsningen heter **Unicode**. Unicode er en tegntabell som pr i dag inneholder mer enn 70 000 tegn, og som potensielt kan inneholde over en million tegn. Unicode dekker snart alle alfabeter som er i bruk i verden i dag, og mange alfabet for utdødde språk, samt mange spesialiserte tegnsett for teknisk eller vitenskapelig bruk. Unicode har siden versjon 3.2 dekket de fleste samiske språk (bortsett fra kildin-samisk), og versjon 4 (som kom sommeren 2003) dekker også kildin-samisk. Alle moderne operativsystem støtter Unicode mer eller mindre (Linux og MacOS X mest, Windows minst). Der støtten ikke er tilfredsstillende i dag, vil den bli bedre i morgen (eventuelt i overimorgen). Unicode er kommet for å bli, og vil mer og mer ta over som standard koding.

## 3. Denne portalen bruker Unicode

Absolutt all tekst og alle nettsider i denne portalen bruker Unicode, og er kodet i UTF-8. Det betyr at du må bruke en nettleser som kan forstå Unicode for å kunne se de samiske tegnene (i tillegg til at du må ha [fonter](#) som inneholder samiske tegn). Dette bør ikke lenger være noe problem, alle nettlesere fra de siste årene støtter Unicode i ulike kodinger. Portalen bruker fonter som enten er innebygde i de vanligste operativsystemene, eller er lette å få tak i (se [siden for fontspesifikasjoner](#) for flere detaljer).

For å kunne søke på samiske ord må du kunne skrive samiske tegn. Det finnes samiske tastatur for både Linux (innebygd i versjoner fra 2003 og senere), MacOS X (innebygd i 10.3) og Windows 2000 (tilgjengelig fra Sametingets [hjemmeside](#)), og kommer snart innebygd i Windows XP. Se referansene for mer informasjon.

For de teknisk interesserte kan nevnes at selv om de bakenforliggende XML-sidene er kodet helt i UTF-8, vil konverteringen til HTML (som skjer automatisk) også konverte en del tegn fra UTF-8-sekvenser til HTML-entiteter. Dette gjelder først og fremst tegn innenfor Latin1 (mer presist gjelder det alle tegn som er definert som entiteter i HTML 4-standard). Dette er en egenskap ved rammeverket som portalen benytter, og vil ikke

endres. Den viktigste konsekvensen av dette er at alle tegn som sendes som entiteter også vil vises korrekt i nettlesere som er så gamle at de ikke skjønner Unicode. Det er derfor mulig å lese de norske sidene, bl.a. denne, og få en forklaring på hvorfor de samiske sidene ikke vises som de skal (svaret på slike problemer er å skaffe en nyere nettleser, men det behøver slett ikke være den nyeste versjonen). Utover dette har det ingen andre praktiske konsekvenser enn at sidene kan bli noen prosent større.

## 4. Referanser

Mange har skrevet mere og bedre om problemene og løsningene enn det vi har gjort her. Nedenfor følger noen referanser:

- [Skolelinux og samisk](#)
- [Trond Trosteruds sider om samisk og data](#)
- [Davvi Girji](#)
- [Svein Lund: MI lille DATABOK](#)
- [Svein Lund: Ingen samisk på Internett?](#)

Felles for alle disse sidene er at de er delvis foreldet, selv om noen av dem ble skrevet så sent som i 2003. Men sidene inneholder uansett mye mere og mer detaljert informasjon om Unicode og koding av samiske tekster enn det det er plass til her. Foreldelsen gjelder først og fremst situasjonen for Unicode i ulike program og operativsystem, som hele tiden blir bedre.