Språkteknologi (SV2122)
Föreläsning 1: Inledning

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Richard Johansson

richard.johansson@svenska.gu.se

22 januari 2014

översikt

praktisk information

kort översikt av ämnet

lagring av text i datorn

nästa föreläsning

schema

- kurshemsida:

  `http://spraakbanken.gu.se/personal/richard/sv2122_2014`
- 10 föreläsningar – observera att salarna varierar (Humanisten eller LT-gatan)!
- 2 datorövningar, båda i T225 (Olof Wijksgatan 6)
- alla tillfällen antingen onsdag eller fredag 10.15–12.00

examination

- redovisning av 2 datorövningar
- tentamen i mars, exakt datum meddelas senare
- betyg: G eller VG, enligt tentamensresultatet

kurslitteratur

- huvudkursbok: Dickinson, M., Brew, C., Meurers, D. *Language and Computers*. Wiley-Blackwell, 2013.
- övrigt: (finns länkade från websidan)
  - Downey, A. *Think Python – How to Think Like a Computer Scientist*.
  - Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*.
  - Wynne, M. (ed). *Developing Linguistic Corpora: a Guide to Good Practice*.
- artiklar länkas från websidan

uppgifter

- två kursuppgifter:
  - enkel programmering i Python (5 februari)
  - användning av Språkbankens annoteringslab (26 februari)
- lös uppgifterna individuellt eller i par, skriv kort rapport
- instruktioner till uppgifterna länkas från hemsidan

översikt

praktisk information

kort översikt av ämnet

lagring av text i datorn

nästa föreläsning

kärt barn har många namn...

- ett tvärvetenskapligt ämne
- "språkteknologi" eller "datalingvistik" – pratar vi om teknologi eller lingvistik?
- på engelska oftast "*natural language processing*" eller "*computational linguistics*"
- en del skiljer på dessa begrepp: det förra skulle vara mer tillämpat och det senare mer grundforskande
- i praktiken väldigt flytande gränser...

betraktat som teknologi

- maskinella metoder för att hantera människors språk
  - "hantera": analysera och generera
  - "språk": skrivet och talat
  - "människors": i motsats till maskiners "språk"
- i den här kursen fokuserar vi på **analys** av **skrivet** språk

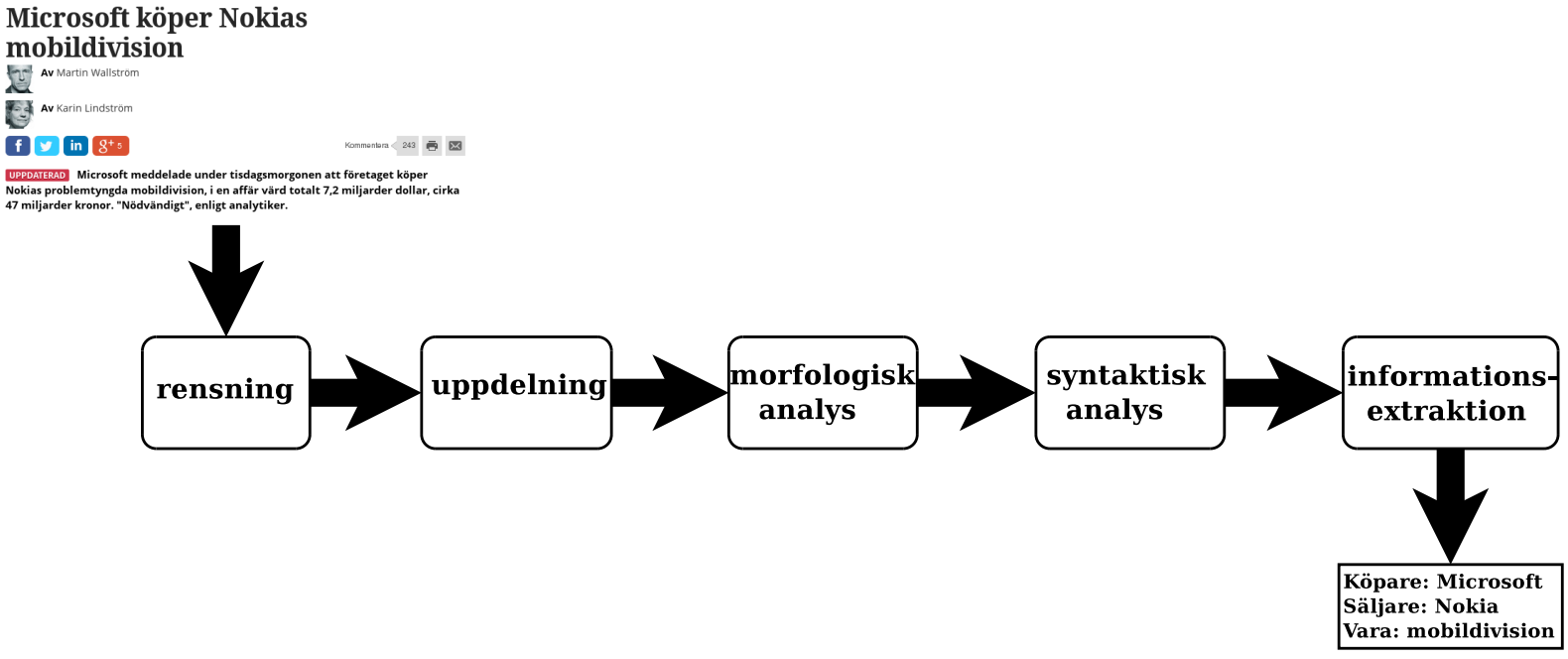
några tillämpningar

- stavnings- och grammatikkontroll, t.ex. i Word
- kategorisering av dokument, t.ex. spamfiltrering
- informationssökning och -extraktion
  - även övervakning av din e-post
- dialogsystem, t.ex. SJs bokning
- talstyrning, diktering
- maskinöversättning, t.ex. Google
- hjälp vid språkinlärning
- stöd för språkvetenskaplig forskning

analysnivåer

- uppdelning av webtext eller ljudinspelning i meningar och ord
- morfologisk analys av ord: “det här ordet (*rev*) är antingen ett verb i preteritum eller ett substantiv i singular, och i detta sammanhang är det ett verb”
- syntaktisk (grammatisk) analys: “den här nominalfrasen är satsens subjekt”
- semantisk analys (betydelse): “den här meningen handlar om att Microsoft (ett företag) köpte en avdelning i Nokia (ett företag)”
- analys av diskurs/argumentation/retorik: “författaren inleder med att peka ut bristerna och därefter föreslå förbättringar”

en typisk språkteknologisk "kedja"



exempel: Språkbankens annoteringslabb

ord	msd	lemma	lex	saldo	prefix	suffix	ref	dephead	deprel
Microsoft	PM. NOM						1	2	SS
köper	VB. PRS. AKT	köpa	köpa..vb.2, köpa..vb.1	köpa..2, köpa..1			2		ROOT
Nokias	PM. GEN						3	4	DT
mobildivision	NN. UTR. SIN. IND. NOM				mobil..av.1, mobil..nn.1	division..nn.1	4	2	OO
.	MAD						5	2	IP

`http://spraakbanken.gu.se/korp/annoteringslabb/`

något om kursens innehåll

utmaningar

- på 1950-talet förväntades maskinöversättningsprogram utvecklas inom några år
  - det blev inte riktigt så (*se ALPAC report*)...
- språkteknologiämnet utvecklas snabbt, men det finns fortfarande begränsingar
- varför är det så? varför är språkteknologi svårare än andra mekaniska uppgifter för datorprogram?
- några orsaker:
  - språkliga enheter kan tolkas på flera olika sätt (mångtydighet, *ambiguity*) beroende på sammanhanget
  - språket är underspecificerat: talaren utelämnar den information som lyssnaren förväntas känna till eller kunna räkna ut själv
  - de lärde tvista...
- svårigheterna finns på alla nivåer av analysen

indelning av texten: svåra fall

- punkter kan orsaka problem

  "...an account with the **U.S. Treasury** to buy Savings Bonds online..."

  "...then I went back to the **U.S. My** dad and I moved..."

- avstavade ord: ska vi ta bort bindestrecket?
- ordindelning i t.ex. kinesiska är inte enkelt att göra automatiskt

exemplet lånat av Liang Huang

utmaning: morfologisk flertydighet

- "*De var jätteduktiga och* ***rev*** *bara ett enda hinder!*"
- "*Jag gick med på att åka ubåt längs ett* ***rev****.*"

utmaning: syntaktisk flertydighet

- "*jag såg en man med ett teleskop*"
- . . . kan läsas på två sätt:
  - *jag såg en man med ett teleskop*
  - *jag såg en man med ett teleskop*

utmaning: lexikal flertydighet

- "*bandet spelar rock*"
  - *band* har 6 betydelser som substantiv enligt lexikonet SALDO
    - snöre
    - för bok
    - grupp
    - samhörighet
    - för ljudinspelning
    - löpande
  - *spela* har 8 betydelser
  - *rock* har 2 betydelser
- problemet kallas **ordbetydelsedisambiguering** (word sense disambiguation)

utmaning: referens, metafor, metonymi, ...

- "*Pelle träffade Nisse i morse innan han åkte till Skåne.*"
  - Vem syftar *han* på?
- "*sjuka som ligger på soffan och odlar sina karaktärsbrister*"
  - Vad betyder *odlar*?
- "*Enligt Moskva iscensattes upproret av tjetjenska rebelledare.*"
  - Vad betyder *Moskva*?

utmaning: textstruktur, "att läsa mellan raderna"

- En recension från Amazon: "*This book was great. I especialy liked the chapter where Bill sticks a falafel in his co-worker's hoo-hah. Priceless*"
- En annan: "*I would very much like to know the names of all the songs contained in this book BEFORE I decide to buy it. How may I access the "contents" page listing all the songs in this book?*"
- Är de positiva eller negativa?

översikt

praktisk information

kort översikt av ämnet

lagring av text i datorn

nästa föreläsning

lagringsmedia

- språkteknologi: program som hanterar text på datorn
- vad betyder “text” och att den “finns på datorn”?
- typiskt att något språkligt material finns i en fil på ett **lagringsmedium**: hårddisk, cd-rom, flashminnen, . . .

filsystem, filer, kataloger, . . .

- ett lagringsmedium innehåller en stor hög med information
- för att göra den hanterbar delar vi in den i **filer**
- . . . och för att hålla ordning på filerna använder vi **kataloger**

representation av information

- det enklaste tänkbara systemet för kommunikation består av två "meddelanden": **på** och **av**



- OBS att systemet i sig inte säger något om hur de två meddelandena ska tolkas (jfr Saussure)
- i datorsammanhang kan detta representeras med en elektronisk "strömbrytare" och kallas då en **bit**
- fysiskt på t.ex. en hårddisk representeras en bit med hjälp av riktningen på ett magnetfält

representation av information, vidare

- mer komplicerad information representeras helt enkelt genom att använda mer än en bit
  - 1 bit: 2 meddelanden
  - 2 bitar: $2 \cdot 2 = 4$ meddelanden
  - 3 bitar: $2 \cdot 2 \cdot 2 = 8$ meddelanden
  - ...
- av sedvana brukar man gruppera bitar i grupper om 8, och dessa brukar kallas **bytes**
  - dvs en byte kan representera 256 olika meddelanden
  - brukar ses som tal mellan 0 och 255

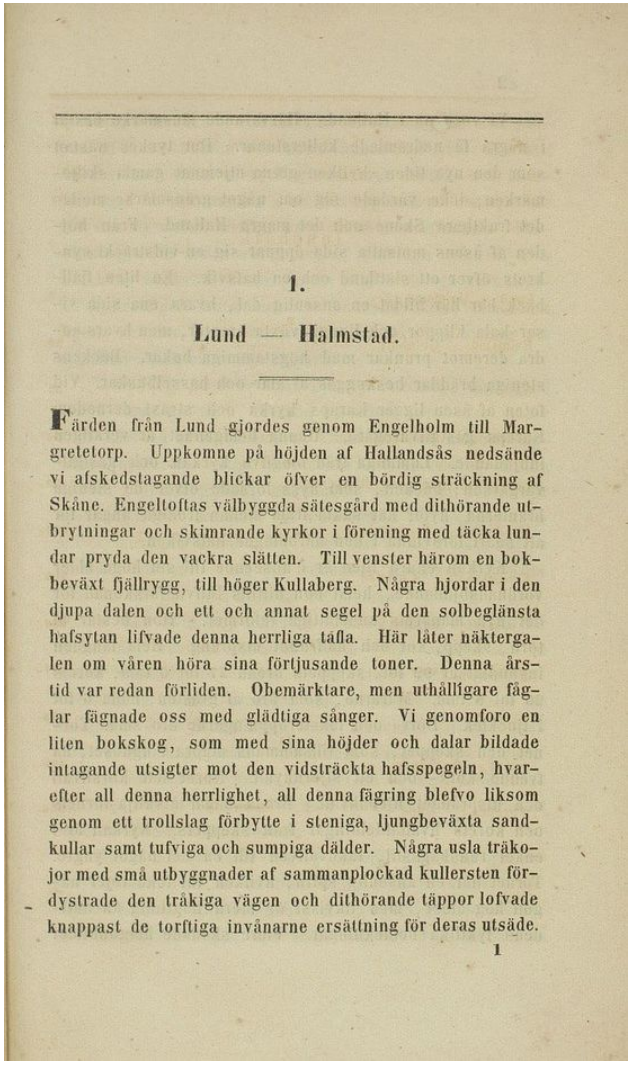
filers anatomi

- namn och plats i katalogstrukturen
- extrainformation t.ex. läs- och skrivrättigheter
- filinnehåll: en hög med bytes
  - filens innehåll meningslöst utan ett program som tolkar det

olika sorters filer med språkligt material

- språkligt material kan finnas i många olika typer av filer
- dessa kan vara uppbyggda på helt olika sätt!
- exempel:
  - ren text: bokstäver och inget annat
  - formaterad text, som innehåller information om teckensnitt och layout, t.ex. Word, PDF, HTML
  - bild av text, t.ex. från en kamera eller scanner
  - inspelning av tal

bild av text kontra textfil



1.

Lund — Halmstad.

Färden från Lund gjordes genom Engelholm till Margretetorp. Uppkomne på höjden af Hallandsås nedsände vi afskedstagande blickar öfver en bördig sträckning af Skåne. Engeltoftas välbyggda sätesgård med dithörande utbrytningar och skimrande kyrkor i förening med täcka lundar pryda den vackra slätten. Till venster härom en bokbeväxt fjällrygg, till höger Kullaberg. Några hjordar i den djupa dalen och ett och annat segel på den solbeglänsta hafsytan lifvade denna herrliga tafla. Här låter näktergalen om våren höra sina förtjusande toner. Denna årstid var redan förliden. Obemärktare, men uthålligare fåglar fägnade oss med glädtiga sånger. Vi genomforo en liten bokskog, som med sina höjder och dalar bildade intagande utsigter mot den vidsträckta hafsspegeln, hvarefter all denna herrlighet, all denna fägring blefvo liksom genom ett trollslag förbytte i steniga, ljungbeväxta sandkullar samt tufviga och sumpiga dälder. Några usla träkojor med små utbyggnader af sammanplockad kullersten fördystrade den tråkiga vägen och dithörande täppor lofvade knappast de torftiga invånarne ersättning för deras utsäde.

1

1.

Lund – Halmstad.

Färden från Lund gjordes genom Engelholm till Margretetorp. Uppkomne på höjden af Hallandsås nedsände vi afskedstagande blickar öfver en bördig sträckning af Skåne. Engeltoftas välbyggda sätesgård med dithörande utbrytningar och skimrande kyrkor i förening med täcka lundar pryda den vackra slätten. . . .

representation och definition av tecken

- datorn lagrar text som en följd av tecken
- tecknen är specificerade i en fix teckenuppsättning
- teckenuppsättningen är numrerad och positionerna kallas *kodpunkter* (*codepoints*)
  - för länge sedan: 128 symboler (ASCII)
  - därefter: 256 symboler (Latin-1, . . .)
  - numera: obegränsat antal symboler (Unicode)
    - `http://www.unicode.org`
- t.ex. texten *lök* motsvaras följande Unicode-symboler:
  - *Latin small letter l*, kodpunkt 108
  - *Latin letter o with diaeresis*, kodpunkt 246
  - *Latin small letter k*, kodpunkt 107

kodning av tecken

- för att spara en text i fil måste vi göra om tecknen till bytes
- denna process kallas **teckenkodning** (*character encoding*)
- kompliceras av att det bara finns 256 möjliga bytes, men betydligt fler möjliga bokstäver ...
- den numera vanligaste kodningen kallas UTF-8
- UTF-8 är en anglocentrisk kodning!
  - engelska bokstäver, samt vanliga skiljetecken → 1 byte per bokstav
  - mer exotiska europeiska tecken → 2 bytes per bokstav
  - kinesiska, japanska, ... → 3 bytes per bokstav
  - hieroglyfer, gotiska, ... → 4 bytes per bokstav
- äldre kodningar t.ex. Latin-1 hade 1 byte per bokstav

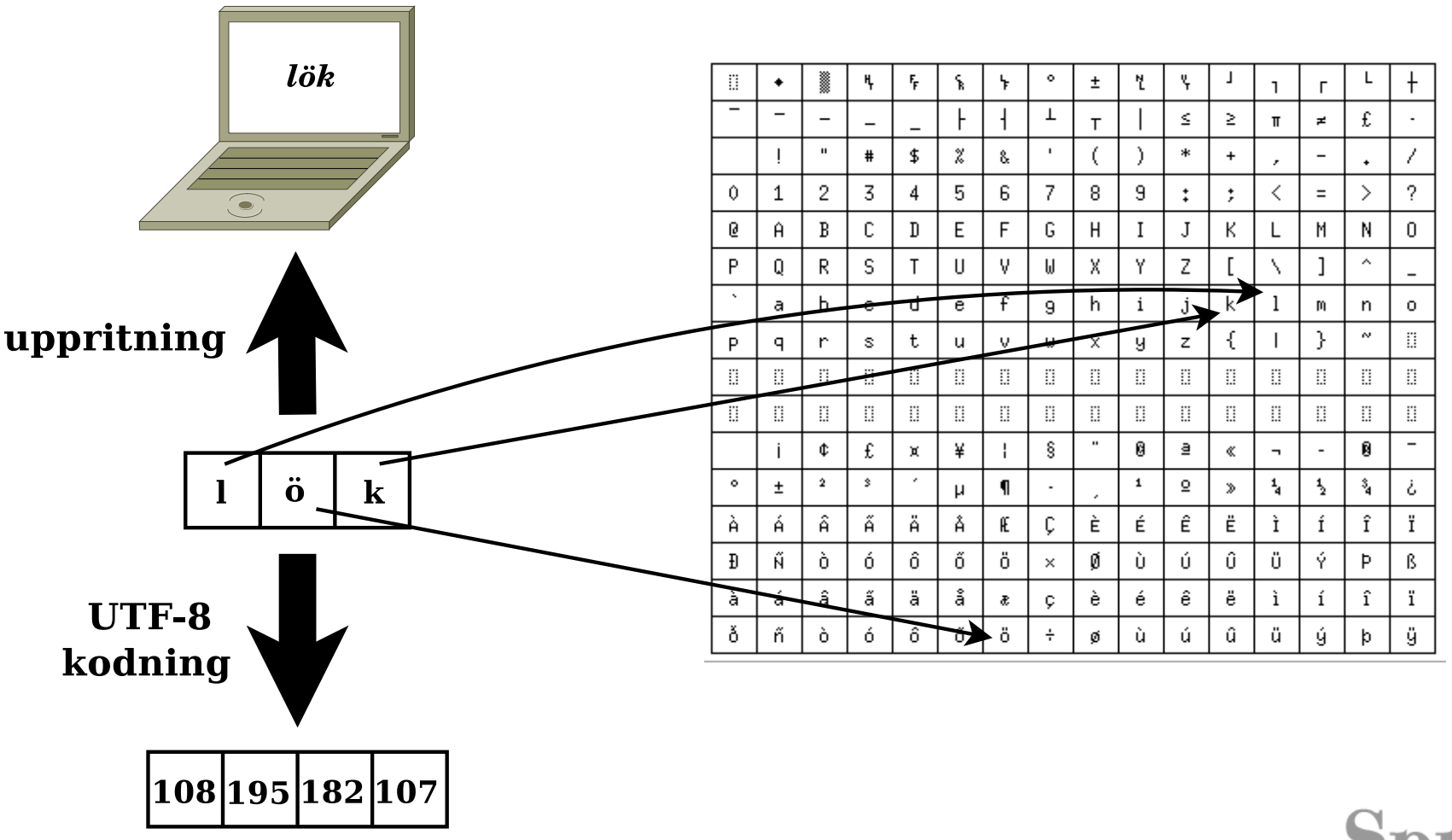
exempel på kodning av Unicode som UTF-8

- *Latin small letter l*, kodpunkt 108 → byte 108
- *Latin letter o with diaeresis*, kodpunkt 246 → bytes 195, 182
- *Latin small letter k*, kodpunkt 107 → byte 107

vad är bokstäver egentligen?

- är ligaturen *fi* en bokstav?
- bokstavsbegreppet ännu mer komplicerat för en del icke-latinska skriftsystem
  - arabiska
  - Devanagari och dess släktingar
  - Hangul
- förutom att definiera tecken så behöver vi en **uppritningsmekanism** (rendering) som kombinerar tecken på rätt sätt

kodning och uppritning . . .

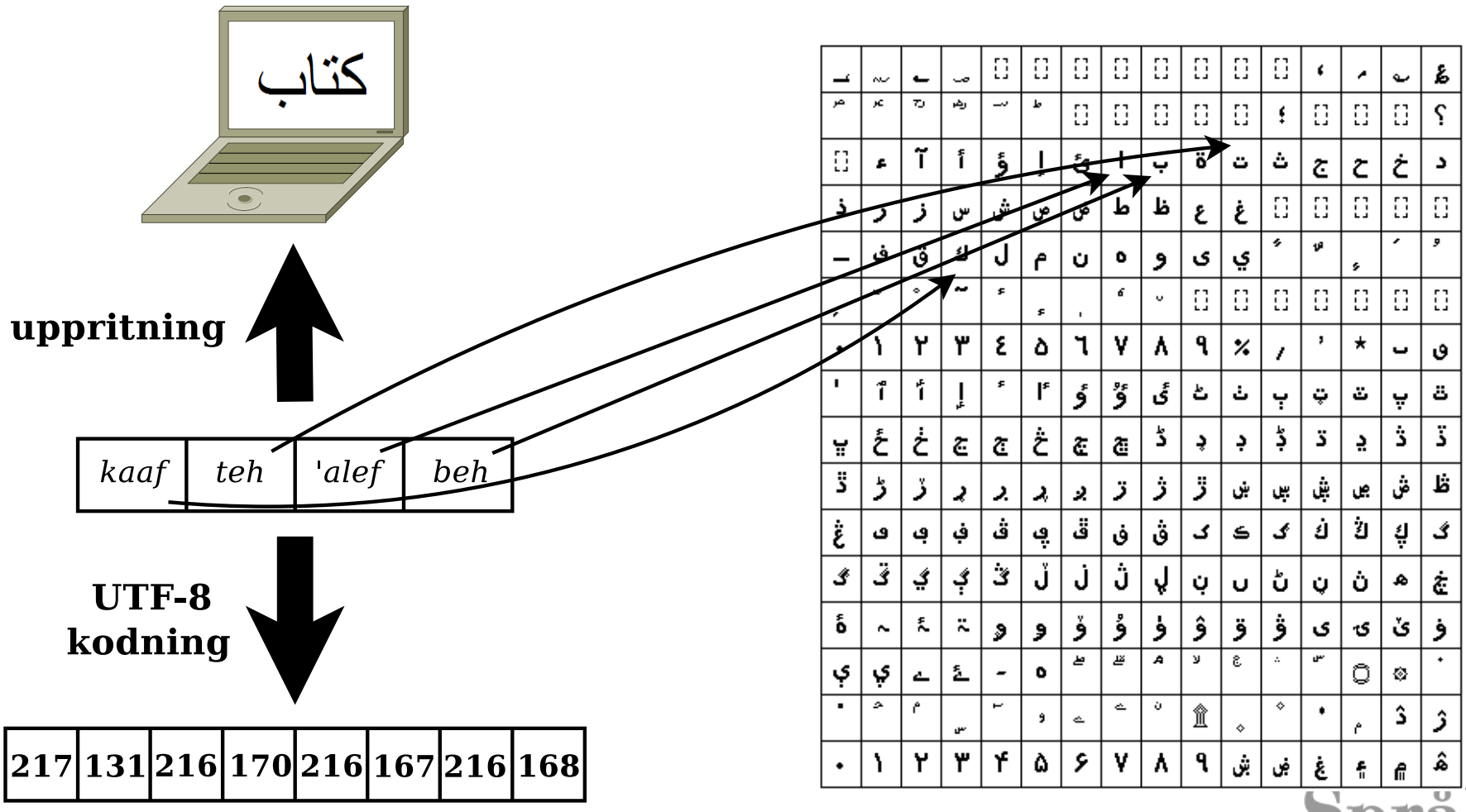


exempel: arabiska

- arabiska bokstäver renderas (ritas) olika beroende på position i ordet
- från höger till vänster!
- men det är ändå samma symboler, och i datorns minne finns inget höger/vänster

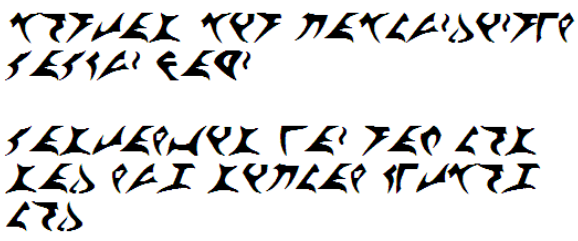
kaf, *teh*, *'alef*, *beh* → كتاب

kodning och uppritning . . .



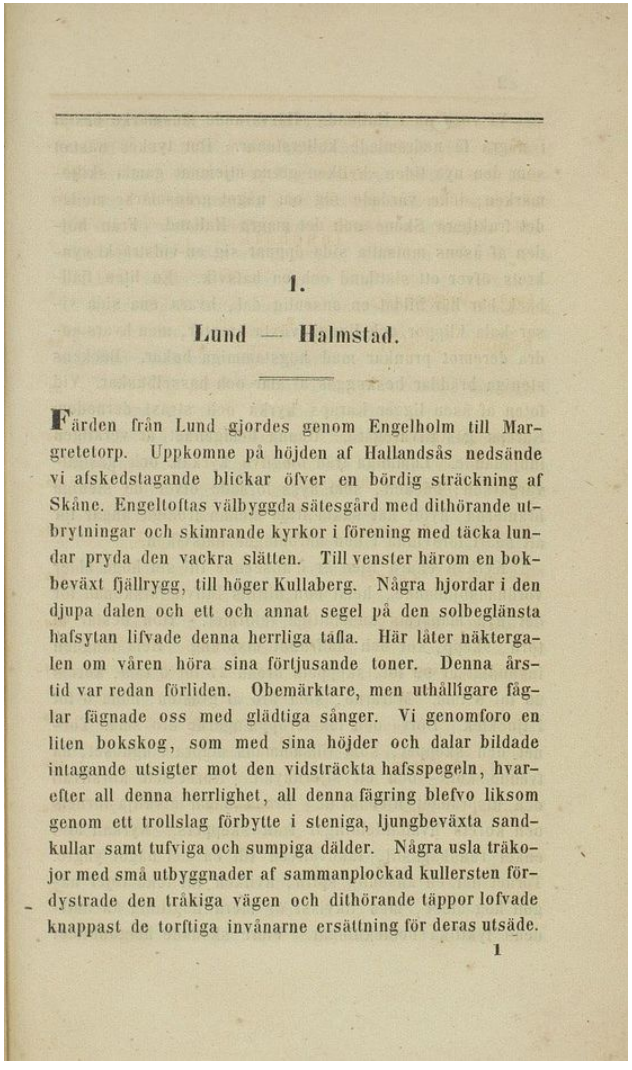
finns det bokstäver till mitt språk?

- `http://www.unicode.org/charts`
- du behöver två saker: dels definition av kodpunkter, och dels teckensnitt eller uppritningsprogram
- de flesta kända skriftsystem (levande och döda) är nu standardiserade i Unicode
- t.ex. äldre, yngre, och kortkvistrunor (`http://www.unicode.org/charts/PDF/U16A0.pdf`) dock varken dalrunor eller hälsingerunor
- ett antal halvstandardiserade system finns också



digitalisering. . .

från bild till text: OCR



1.

Lund — Halmstad.

Färden från Lund gjordes genom Engelholm till Margretetorp. Uppkomne på höjden af Hallandsås nedsände vi afskedstagande blickar öfver en bördig sträckning af Skåne. Engeltoftas välbyggda sätesgård med dithörande utbrytningar och skimrande kyrkor i förening med täcka lundar pryda den vackra slätten. Till venster härom en bokbeväxt fjällrygg, till höger Kullaberg. Några hjordar i den djupa dalen och ett och annat segel på den solbeglänsta hafsytan lifvade denna herrliga tafla. Här låter näktergalen om våren höra sina förtjusande toner. Denna årstid var redan förliden. Obemärktare, men uthålligare fåglar fägnade oss med glädtiga sånger. Vi genomforo en liten bokskog, som med sina höjder och dalar bildade intagande utsigter mot den vidsträckta hafsspegeln, hvarefter all denna herrlighet, all denna fägring blefvo liksom genom ett trollslag förbytte i steniga, ljungbeväxta sandkullar samt tufviga och sumpiga dälder. Några usla träkojor med små utbyggnader af sammanplockad kullersten fördystrade den tråkiga vägen och dithörande täppor lofvade knappast de torftiga invånarne ersättning för deras utsäde.

1

⇒

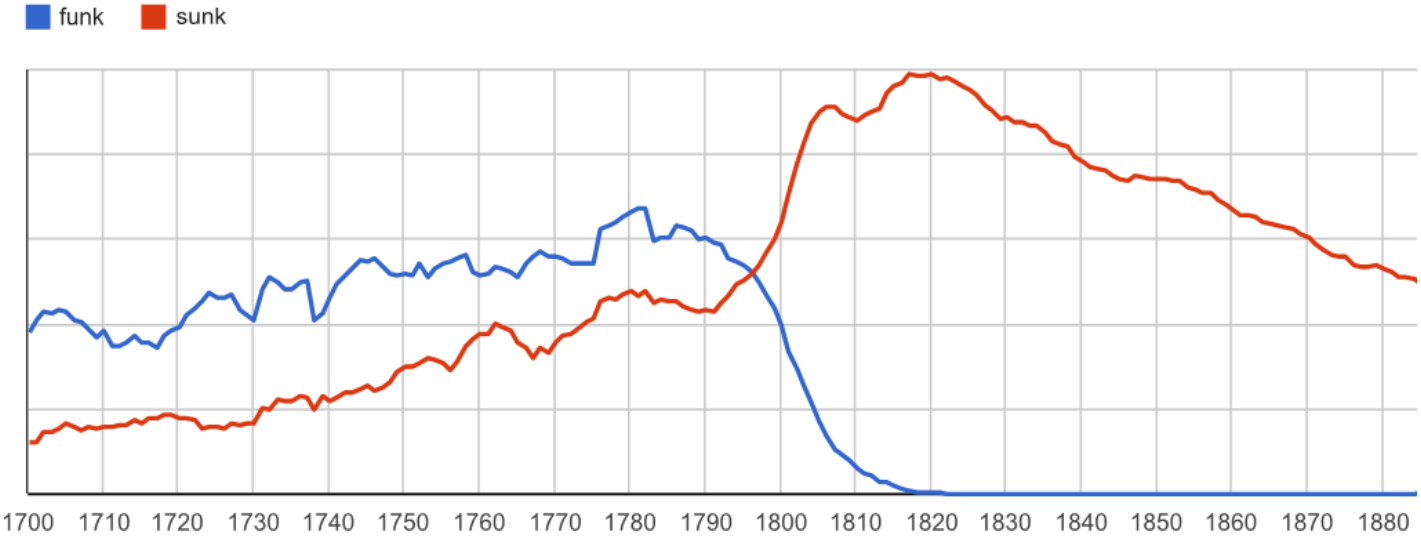
1.

Lund – Halmstad.

Färden från Lund gjordes genom Engelholm till Margretetorp. Uppkomne på höjden af Hallandsås nedsände vi afskedstagande blickar öfver en bördig sträckning af Skåne. Engeltoftas välbyggda sätesgård med dithörande utbrytningar och skimrande kyrkor i förening med täcka lundar pryda den vackra slätten. . . .

när OCR ställer till det

- Google ngram viewer: `http://books.google.com/ngrams`



`http://languagelog.ldc.upenn.edu/nll/?p=2848`

översikt

praktisk information

kort översikt av ämnet

lagring av text i datorn

nästa föreläsning

nästa föreläsning

- en mycket översiktlig inledning till datorprogrammering
- vi kommer att använda programspråket Python och språkteknologiverktyget NLTK
- detta är också temat för första datorövningen
- på fredag i samma sal!