CLT131 Korpusten käsittely
490160-0
Viides luento

Nicholas Volk

Yleisen kielitieteen laitos, Helsingin yliopisto

HELSINGIN YLIOPISTON
YLEISEN KIELITIETEEN LAITOS

Syötteen lukemisesta

- Aikaisemmin Perl-käskymme ovat lukeneet syötettä rivin kerrallaan.
- Muuttujaan `$/` on talletettu kerralla luettavan alueen erotin.
- Oletusarvoisesti `$/ = "\n"`
- Arvoa voi muuttaa tai sen voi asettaa pois päältä käskyllä `undef $/;`
- `$/`:n ollessa määrittelemätön, luetaan kerralla koko syöte
- `$/` vastaa vastaa kutakuinkin Awk:n RS-muuttujaa

BEGIN-lohko

- `BEGIN`-lohko suoritettaan ohjelman aluksi.
- Kun komennot ovat muotoa

  `perl -pe 'BEGIN { undef $/; } käskyt; '`

  luetaan rivin sijasta koko syöte kerralla muistiin.
- Tavuviivojen poistoa:

```
perl -pe 'BEGIN { undef $/; }
s/([a-zåäö])-\n([a-zåäö])/$1$2/g; '
```

- Kovin Awk-maista Perliä...

Lisää s///:n optiota

- Optio `s` tulkitsee merkkijonon yksiriviseksi, eli `\n` on merkki muiden joukossa
- Optio `m` on tavallaan sen vastakohta, merkkijono tulkitaan useita rivejä sisältäväksi, eli `^` ja `$` voivat löytyä monta kertaa.

```
$ perl -e '$_="abc\n"; s/^.*$//; print $_;'

$ perl -e '$_="abc\n"; s/^.*$//m; print $_;'

$ perl -e '$_="abc\n"; s/^.*$//s; print $_;'
$
```

Lisää esimerkkejä

```
$ perl -e '$_="abc\ndef\n"; s/a.*e//; print $_;'
abc
def
$ perl -e '$_="abc\ndef\n"; s/a.*e//m; print $_;'
abc
def
$ perl -e '$_="abc\ndef\n"; s/a.*e//s; print $_;'
f
$ # alusta loppuun:
$ perl -e '$_="abc\ndef\n"; s/^.*$//; print $_;'
abc
def
$ perl -e '$_="abc\ndef\n"; s/^.*$//m; print $_;'

def
$ perl -e '$_="abc\ndef\n"; s/^.*$//mg; print $_;'


$ perl -e '$_="abc\ndef\n"; s/^.*$//s; print $_;'
$
```

Annotoitu korpus

- Korpus on siis kärjistetysti iso kasa tekstiä
- Annotoituun korpukseen on lisätty lisäksi lingvististä tietoa
- Brownin korpuksessa annotointi merkittiin sanan perään:

```
to_TO
stop_VB
Mr_NPT
Gaitskell_NP
from_IN
...
```

- Tulkintavaihtoehtojen määrä pieni englannin kielessä
- Suomessa substantiiveilla on n. 2200 tulkintaa ja verbeillä n. 12000 (Karlsson 1982)

Rakenteinen dokumentti

- Raakatekstilläkin on yleensä selkeä rakenne
- Otsikkoa seuraa yksi tai useampi lauseista koostuva kappale
- Rakenteisessa dokumentissa rakenne (korpuksessa saneet) ja sisältö (esim. syntaktiset suhteet) on eriytetty toisistaan
- Rakenteen kuvaamisen on useita eri kieliä, mm. LaTeX- ja XML-pohjaiset kielet
- Mm. kielipankin aineistot XML-muodossa (osa vielä SGML:nä)
- Opettelemme tunnistamaan XML-taggauksen ja irroittamaan sisällön rakenteesta
- Rakenteisille dokumenteille on laitoksella oma kurssinsa

Äskeisen kalvon rakenne

```
\begin{slide}{Rakenteinen dokumentti}
  \small
  \medskip
  \begin{itemize}
  \item Raakatekstilläkin on yleensä selkeä rakenne
  \item Otsikkoa seuraa yksi tai useampi lauseista koostuva kappale
  \item Rakenteisessa dokumentissa rakenne (korpuksessa saneet)
        ja sisältö (esim. syntaktiset suhteet) on eriytetty toisistaan
  \item Rakenteen kuvaamisen on useita eri kieliä,
        mm. \LaTeX- ja XML-pohjaiset kielet
  \item Mm. kielipankin aineistot XML-muodossa (osa vielä SGML:nä)
  \item Opettelemme tunnistamaan XML-taggauksen ja irroittamaan
        sisällön rakenteesta
  \item Rakenteisille dokumenteille on laitoksella oma kurssinsa

  \end{itemize}
\end{slide}
```

XML-dokumentin rakenne

- XML-dokumentin rakenne on määritetty Document Type Definitionissa
- DTD voi olla erillisessä tiedostossa tai XML-dokumentin alussa
- DTD kertoo mitä tageja (merkitsimiä) dokumentti saa sisältää
- DTD ei kosketa meitä tällä kurssilla, joten oletetaan sen olevan erillään
- Dokumentin määritelmänmukaisuus (well-formedness) ei sekään kiinnosta meitä

XML-tagit

- Tagit alkavat <-merkistä ja päättyvät >-merkkiin
- Alkutagin <:n perässä tulee tagin nimi: `<otsikko>`
- Lopputagi on muotoa `</otsikko>`
- Tagi, joka ei vaadi lopputagia merkitään laittamalla >:n eteen: `<rivinvaihto/>`
- Alkutagi voi saada tarkentavia atribuutteja:

```
<otsikko koko="iso" >
<laatikko leveys="200" korkeus=50>
```

- Eli numeerisia arvoja ei ole pakko laittaa lainausmerkkien sisään...
- Mutta se on hyvä tapa...

Erikoismerkit

- Tageissa käytettyjä <- ja >-merkkejä ei saisi kirjoittaa dokumenttiin sellaisenaan
- Niille on olemassa oma notaationsa, jossa &-merkin ja puolipisteen väliin kirjoitetaan merkin "tunnus":
  < on `<` ja > on `>`
- Vastaavasti &-merkki pitää kirjoittaa poikkeuksellisesti: `&`
- Kaikkia merkkejä ei saa esiin näppäimistöltä, niillekin voi olla omat erikoisnotaationsa: ä on `&auml;`, Ö `&Ouml;`...
- Lisäksi merkkeihin voi viitata numeeristen tunnusten avulla: esim. `&#38;`
- Tämä taas vain tiedoksi, ei oikeastaan tämän kurssin asiaa...

Kommentit

- Kommentti alkaa merkkijonolla `<!--`
- Kommentti päättyy merkkijonoon `-->`
- Kommenttiin kuuluva alue jätetään huomioitta.
- Kommentit kannattaa heittää hiiteen jo ennen kun muuta dokumenttia aletaan käymään läpi...
- Kommentin alku ja loppu ei aina ole samalla rivillä, kuten ei taginkaan.

Erot SGML:ään

- SGML on XML:n edeltävä, metakielissä on pieniä eroja
- Lopputagiton tag ei tarvitse loppumerkintää

  `<rivinvaihto>`
- Kirjainkoolla ei ole väliä SGML:ssä:

  `<HTML>`

  `<html>`

  `<hTmL>`

Merkkien esittäminen tietokoneella

- Kaikkia merkkejä ei aina pysty syöttämään näppäimistöltä
- Merkeillä on käytetyn merkkistandarin (mm. ISO-8859-1 ja Unicode) mukaiset numeeriset arvot
- ISO-8859-perheen 128 ensimmäistä merkkiä ovat kaikissa samat (7-bittinen ASCII-merkistö), loput 128 merkkiä ovat sisällöltään erilaisia.
- Esim. ISO-8859-1 sisältää meille rakkaat ääkköset.
- Aakkoset menevät (käytännössä) aina oikein. Ääkköset näkynevät väärin eri merkkistandardia käytettäessä.

Oktaali- ja heksaluvut

- Käskyn `tr ' ' '\012'` osa `\012` on oktaalikuku, joka tarkoittaa *Line Feed* -näppäintä eli Unixin rivinvaihtomerkkiä.
- Esim. `\007` on piip-ääni ainakin ASCII-pohjaisissa järjestelmistä.
- Perlissä voi viitata merkkiin myös heksalukuna (16-järjestelmä)
- Tällöin kenoviivan perään laitetaan `x` ja kaksinumeroinen heksaluku (`[0-9A-F]`):

```
$ perl -pe 'tr/BA/\x41\102/;'
AB
BA
$
```

Susanne-korpuksen rakenne

http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/public/susanne.html

```
G01:0010b    JJ      NORTHERN     northern    [O[S[Np:s.
G01:0010c    NN2     liberals     liberal     .Np:s]
G01:0010d    VBR     are          be          [Vab.Vab]
G01:0010e    AT      the          the         [Np:e.
G01:0010f    JB      chief        chief       .
G01:0010g    NN2     supporters   supporter   .
G01:0010h    IO      of           of          [Po.
G01:0010i    JJ      civil        civil       [Np.
G01:0010j    NN2     rights       right       .Np]
```

Susanne-korpuksen rakenne

```
G01:0010b    JJ      NORTHERN     northern    [O[S[Np:s.
```

- Kentät ovat sarkain-merkein (`\t`) erotetut.
- Field 1: text references
- Field 2: Part of speech tags (morfologinen tulkinta)
- Field 3: The text words (saneet)
- Field 4: Base form (lemma)
- Field 5: Syntactic annotation
- `[` viittaa jonkin syntaktisen kokonaisuuden alkuun, `]` vastaavasti loppuun

Connexorin FDG-jäsentimen tulostusta

- Connexorinkin jäsentimen antama tekstituloste on selkeä:
  1. Saneen sijainti lauseessa
  2. Sane
  3. Lemma
  4. Funktionaalinen dependenssi
  5. Pintasyntaktinen tagi ja morfologinen analyysi
- Kentät on erotettu sarkainmerkein
- Kenttä 4 voi olla myös tyhjä!
- Kenttä 5 voi olla moniselitteinen, jokaisella tulkinnalla oma kenttä!
- Eli kutakin sanetta kohden on 5+ kenttää! (Yleensä 5)

Kollokaatio

- Kollokaatio tarkoittaa kahden tai useamman sanan tavallista suurempaa yhdessäesiintymistodennäköisyyttä
- Muuttamalla frekvenssilista absoluuttisesta suhteelliseksi saadaan sanojen esiintymistiheydet
- Kollokaatio viitannee jonkinlaiseen semanttiseen suhteeseen
- Tarkasteltavien sanojen tarvitsee olla yleisiä, jotta niiden kollokaatiosta voidaan sanoa jotakin
- Helppo tilastoida, vaikea tulkita