CLT131 Korpusten käsittely
Kuudes luento, 13.12.2005

Nicholas Volk

Yleisen kielitieteen laitos, Helsingin yliopisto

HELSINGIN YLIOPISTON
YLEISEN KIELITIETEEN LAITOS

Puhekorpukset

- Puhetta voidaan esittää myös tekstinä (esim. keskustelunanalyysi)
- Ihanteellisesti alkuperäinen puhe (äänisignaali) on nauhoitettu ja tallennettu digitaalisesti.
- Äänisignaaliin voidaan yhdistää tekstimuotoista tietoa esim. äänne-, sana- ja lausekerajoista.

Keskustelunanalyysi

- `http://www.helsinki.fi/hum/skl/ca/`
- Altis subjektiivisuudelle, en ole nähnyt esim. alkukahdennusta merkittävän koskaan.
- Käytetyn notaation (ks. yllä oleva linkki) saaminen esittäminen raakatekstinä hankalaa. `http://childes.psy.cmu.edu/manuals/CHAT.pdf`

Puheen visualisointi - Praat

- Praat (`http://www.praat.org`) on erittäin monipuolinen
- Vastaavasti laajuutensa takia aluksi vaikea hahmottaa
- Puheen äänne- tms. rajat ym. informaatio voidaan merkitä ja tallettaa TextGrid-nimiseen tietorakenteeseen.
- Mietta Lennes on pitänyt Praat-kursseja laitoksella
- Mahdollistaa omien skriptien käytön
- Must foneetikoille

Puheen visualisointi - Wavesurfer

- Wavesurfer (`http://www.speech.kth.se/wavesurfer/`) on kevyempi ja intuitiivisempi
- Alunperin niin intuitiivinen, ettei sille ole kirjoitettu käyttöohjetta, vaikka ohjelma onkin ajan myötä monipuolistunut...
- Tämän päälle on rakennettu mm. CSC:n Puh-editori `http://www.csc.fi/kielipankki/puhe/`
- CSC:ltä löytyy myös puheaineistoja:

  `http://www.csc.fi/kielipankki/aineistot/puhe.phtml`

Wavesurferin käyttö

- Ei toimi palvelimilta (HY:n venus), koska etäyhteyden palvelin ei voi tunkea ääntä paikallisen koneen äänikorttiin.
- `wavesurfer` *`äänitiedosto`* (tiedoston voi toki valita ohjelman sisältäkin).
- Choose configuration -valikosta meille sopii kohta *Speech analysis*.
- Se avaa kuvan aaltomuodosta, spektogrammin ja näkymyksensä puheen sävelkulusta.

Visualisointi ja mitä siitä näkee

- Spektri, sointi, formantti
- Aaltomuoto, amplitudi, äänihuuliperiodi
- Perustaajuus, yläsävel

Ääniformaatit

- Tiedostopäätteitä: au, ogg, mp3, raw, snd, wav (riff) ja lukemattomia muita.
- Äänidata vie paljon levytilaa, joten osa formaateista pakkaa ääntä usein tietoa hävittäen, esim. Mp3.
- WAV yleisin, siitä eri muotoja, jotka kerrotaan tiedoston otsakkeessa (header)
- Ihmiskorva ei kuitenkaan välttämättä kuule eroa.
- Samaten näytteenottotaajuus voi vaihdella (CD-levy: 44100 näytettä).
- Samaten äänelle voi olla monta eri kanavaa (mono, stereo...)

Ääniformaatit (2)

- Näytteenottotaajuus/2 on korkein tallentuva hertsimäärä, CD:llä siis 22050.
- Ihmiskorva ei kuitenkaan yleensä kuule läheskään näin korkeita ääniä.
- Lankapuhelimen kaista n. välillä 300-3000 hertsiä, joten puhelimeen on turha soittaa mitään 6000:tta näytettä sekunnissa suurempaa.
- Puhelimessa käytetään mm. μ-LAW (eli u-law tai mu-law) ja A-LAW -muotoista. Nämä ovat WAV/RIFF-formaatin kompressoituja muotoja.

Konversio

- kooderi, dekooderi
- lame, sox ...
- resamplaus: ch_wave, sox...

Lemmie

- Mickel Grönroosin laitoksella ja CSC:llä kehittämä työkalu
- WWW-demo