Kõnekeele ressursid ja kõnetehnoloogia andmebaasid

Einar Meister

TTÜ Küberneetika Instituut

Eesmärgid

Eesti keele foneetilisteks ja kõnetehnoloogilisteks uuringuteks ning arendustöödeks vajalike kõnekorpuste salvestamine, digitaliseerimine, märgendamine ja arhiveerimine, samuti ühtse tehnoloogilise keskkonna loomine erinevate andmebaaside haldamiseks ja efektiivseks kasutamiseks.

Projekti alateemad:

- Uudistekorpus
- Loengukõne korpus
- Aktsendikorpus
- Infrastruktuur

Uudistekorpus

- *Ca* 300 tunni Eesti Raadio lühiuudiste salvestusi ajavahemikust 29.11.2005- 29.05.2006 (10 erinevat raadiodiktorit)
- Digitaliseeritud (skaneeritud ja teisendatud OCR programmiga tekstiks) ca 10000 lk uudistetekste
- Märgendamine:
  - Esmane lähenemine: tekstifailide ja salvestuste käsitsi ühestamine – ei tööta!
  - Innovatiivne lähenemine: (1) uudistekstide põhjal luuakse kõnetuvastuse keelemudel ja salvestuste transkriptsioon saadakse kõnetuvastaja abil (T.Alumäe projekt)

    (2) tekstide kontroll ja salvestustega ühestamine käsitsi

Märgendamise töökeskkond

- ***Transcriber*** (http://trans.sourceforge.net/)
- Signaali- ja transkriptsioonifailid asuvad labori serveris, märgendajatel on turvaline ligipääs üle Interneti
- Kontrollitud ca 20 tundi kõnematerjali transkriptsioonid

Märgendamise reeglid

- Põhiüksuseks on uudislugu, so ühte sündmust kajastav saatelõik
- Uudislugu jagatakse lauseteks sisu ja prosoodilise informatsiooni alusel
- Lause lõpu tähistamiseks kasutatakse punkti, osalausete eristamiseks koma
- Laused ja nimed kirjutatakse suure algustähega, lühendid (näiteks NATO jms) kirjutatakse suurte tähtedega
- Numbrid ja aastaarvud kirjutatakse lahti sõnadena
- Erinevad diktorid tähistatakse vastavate koodidega
- Mürad, muusika ja võõrkeelsed sõnad eristatakse vastavate märgenditega

Aktsendikorpus

- Sisaldab eesti keelt võõrkeelena kõnelevate inimeste kõnesalvestusi
- On vajalik:
  - aktsendinähtude akustilis-foneetiliseks uurimiseks
    → rakendus keeleõppes (metoodika, hääldustreening)
  - aktsendiga kõne akustiliste mudelite treenimiseks
    → rakendus kõnetuvastuses

Aktsendikorpuse kavand

- Ca 20 eri emakeelega keelejuhi eestikeelse kõne salvestused:
  - vene ja soome keel – ca 30 kõnelejat
  - EL “suuremad” keeled (inglise, saksa, prantsuse, hispaania) – ca 15 kõnelejat
  - EL “väiksemad” keeled – 4-6 kõnelejat
  - muud keeled – 1-2 kõnelejat
- Laboratoorne kõne – põhiliselt teksti lugemine, vähesel määral spontaanne kõne (enesetutvustus, piltide kirjeldamine)
- Igalt keelejuhilt ca 20 minutit kõnet

Tekstikorpus

Sisaldab olulisemaid eesti keelele omaseid fonoloogilisi nähtusi:

- Vokaalid, diftongid, konsonandid ja konsonantühendid
- Palatalisatsioon
- Sõnaprosoodia, välted
- Lauseprosoodia

Näiteid tekstikorpusest

Kaotasin sada krooni raha.
Palun saada talle artikli koopia.
Soovin saada kolme kuu aruannet.

Sügisel pandi vili salvedesse.
Selle viili käepide oli murdunud.
Uut viili kasutage ettevaatlikult.

Mees ajas kiusu ainult kiusu pärast.
Ära veereta päeva päeva järel tegevusetult õhtusse.

Äi äigas peoga üle näo ja tegi lõo ning käo häälitsusi.
Jõululaupäeva õhtul sõime verivorsti ja sülti.

Mehe käitumine tegi talle haiget.
Ema pani lapsele teki peale.
Voodil vigises tekki mähitud beebi.

Kamina võre tehti rauast.
Antud võrre ei sobi ülesande lahendamiseks.
Vanu võrre ta ei paranda.

Paigutasin oma raha kinnisvarasse.
Firma oli üsna krahhi piiril.
On oodata krahhi panganduses.

Looma kõrva taha tehti süst.
Haiget kõrva soojendage viis päeva.

Ma ei mõista seda keelt.
Et mõista kaaslast, on vaja ta ära kuulata.

Mul on üks väike palve Tiinale.
Pole lootustki, et ta selle palve täidab.

Kas korstnapühkija tuleb homme või ülehomme?

Kuidas on selle pikkade juustega tüdruku nimi?

Kas sa tead Peetri uut telefoninumbrit?

Kas te soovite teed või kohvi?

Kirjelda pilti (1)

Kirjelda pilti (2)

Kirjelda pilti (3)

Keelejuhid

- Valikukriteeriumid:
  - emakeel ei ole eesti keel
  - kõnes peab olema kuuldav võõrkeele aktsent
  - võimeline lugema eestikeelset teksti
  - ei esine kõne- ja kuulmishäireid
- Keelejuhi ankeet:
  - vanus
  - emakeel
  - kodune keel,
  - võõrkeelte oskus
  - eesti keele õppimise alustamise iga
  - kui sageli kasutab eesti keelt
  - jm

Keelejuhtide värbamine

- Värbamine:
  - Kolleegid, tuttavad, keeleõpetajad (TLÜ, TTÜ), kõrgkoolide välistudengid, vahetusõpilased (YFU Eesti) jne
  - "Lumepalli" meetod – keelejuht värbab uusi inimesi
  - Kontaktid välisülikoolidega – Helsingi, Turu, Oulu, Viin, INALCO (Pariis), Uppsala, Riia, Vilnius
  - Kontaktid saatkondadega, erialaste ühendustega, rahvusvaheliste firmadega

- **Keelejuhid saavad tasu!**
- **Keelejuhtide leidmine ON raske ülesanne!**

Salvestused

Hetkeseis – 135 salvestust

- **K2 keelejuhid:**
  - vene – 40
  - soome – 30
  - prantsuse - 12
  - saksa – 13
  - itaalia – 4
  - hispaania – 2
  - taani – 2
  - hollandi – 2
  - slovaki – 2,
  - inglise – 2
  - šoti – 1
  - iiri – 1
  - aserbaidžaani – 1
  - jaapani – 1
  - läti – 1
  - leedu – 1
- K1 keelejuhid: 20

Näiteid salvestustest – laused

- Lapse rõõmu pärast olen valmis paljuks.

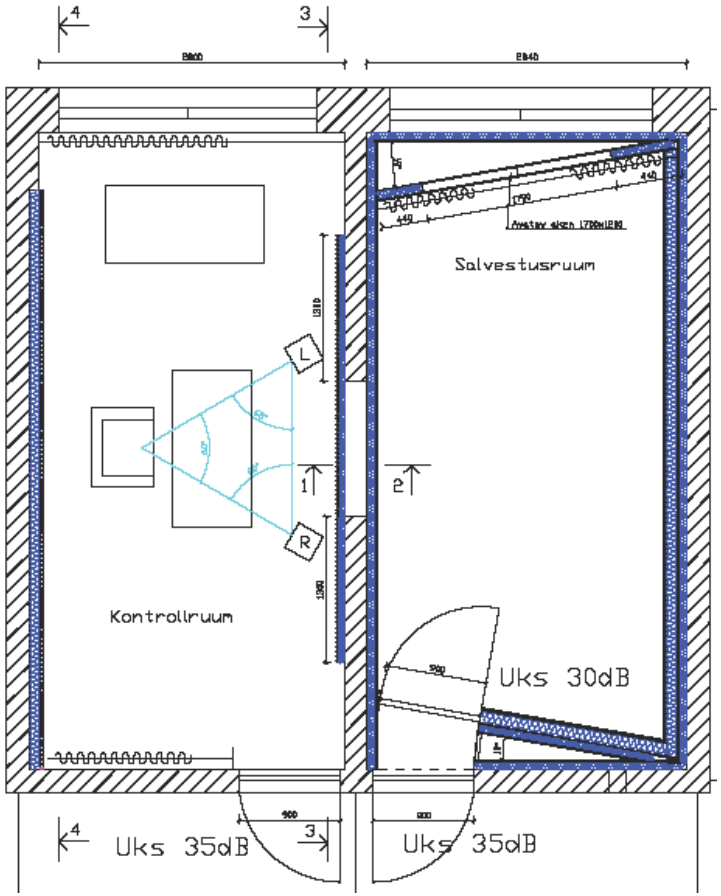
- Kass püüdis muti kinni.

- Kas sa tead Peetri uut telefoninumbrit?

Loengukõne korpus

- üle 100 tunni eri ainevaldkondade akadeemiliste loengute salvestustusi (erinevate lektorite arv on 15)
- ca 4 tundi konverentsiettekandeid (20 isikut, keskmine ettekande pikkus 20 min).
- korpus on ettevalmistamisel märgendamiseks Transcriber'iga.

Infrastruktuur: salvestusstuudio



Infrastruktuur: salvestustehnika

- Mikrofonid:
  - AKG, Sennheiser, Behringer
- Kõlarid:
  - Genelec. JBL
- Mikserpult Mackie Onyx 1640, FireWire I/O-kaart
- Salvestustarkvara Adobe Audition 3.0, SpeechRecorder
- Mobiilsed salvestusvahendid:
  - M-Audio MicroTrack 24/96
  - Edirol R1
  - Digigram VXpocket 440
- Videokaamera Sony HDR-SR12E
- Arvuti Mac Pro + 30” Apple Cinema HD monitor
- Audio/videotöötlus tarkvara Cubase 4, Final Cut Express
- Välised kõvakettad Lacie 2d quadra 2 x 750 GB

Infrastruktuur

- Kõnekorpuste server Dell PowerEdge R200, kõvaketta maht 2TB
- Korpuste haldamiseks ja kasutamiseks kohandatakse korpuste haldussüsteem **LAMUS** – Language Archive Management and Upload System (Max Planck'i Psühholingvistika Instituudis) http://www.lat-mpi.eu/tools/lamus/

Edasine töö

- **Uudistekorpus** – jätkub transkriptsioonide kontroll (ca 100 tundi 2009 lõpuks)
- **Aktsendikorpus** – salvestused 50-60 keelejuhiga

Otsime uusi keelejuhte!

- **Loengukõne korpus:**
  - loengusalvestused TTÜs, ca 50 tundi
  - konverentsiettekannete salvestused, ca 20 keelejuhti
  - loengukõne märgendamine, orienteeruv maht 10 tundi

Edasine töö

- **Jutusaadete korpus** – jutusaadete salvestused (näiteks Kuku-raadio saated Nädala tegija, Vastasseis, Pressiklubi, Välismääraja jms)
  - salvestuste märgendamine ca 10 tundi
- **Häälkäskluste korpus** – salvestus 20 keelejuhiga (tingimusel, et standardi töörühm avaldab eestikeelsete häälkäskluste loendi)
- **Infrastruktuuri arendus:**
  - korpuste haldussüsteemi LAMUS arendus
  - ligipääs kõigile korpustele LAMUSe kaudu (2010)

Tegijad, rahastamine

Põhitäitjad:

Einar Meister
Lya Meister
Rainer Metsvahi
Lepingulised töötajad (uudiste märgendus, loengusalvestused)

Finantseerimine:

2006 – 660 000 EEK
2007 – 400 000 EEK
2008 – 750 000 EEK, sh soetused 150 000 EEK
2009 – 630 000 EEK