Oppitunti 6: Harjoitukset

Oppimistavoitteet

- Annotaation (esim. litteroinnin) aloittaminen, tallentaminen ja annotaatiotyön jatkaminen ELANilla
- Litterointitekniikka ELANilla
- Keskustelun litteraatin vienti ELANista tekstitiedostoon
- Annotaatiokerrosten *tyypittäminen* ELANissa
- Litteroidun kerroksen saneistaminen eli *tokenisointi* ELANissa
- Hakujen tekeminen ELANilla annotoidusta aineistosta

Tehtävät

1. Mikäli koneellasi ei vielä ole ELAN-ohjelmaa, lataa ja asenna se (https://archive.mpi.nl/tla/elan).

   Huom. Mikäli käytät yliopiston koneita, saatat joutua pyytämään IT-tukihenkilöäsi asentamaan ohjelman. Helsingin yliopiston keskitetyssä ylläpidossa olevilla koneilla ELANin pitäisi löytyä myös yliopiston tarjoamana valmiina asennuspakettina suoraan Software Centeristä (ks. ohje Windows-koneille).

2. Avaa ELAN-ohjelma valmiiksi.

3. Katso ELAN-ohjelmalla litterointia käsittelevä video: https://aoe.fi/#/materiaali/1637

   Halutessasi voit koettaa videon katsottuasi tehdä vastaavat asiat perässä samalla aineistolla. Videolla käytetään vapaasti saatavilla olevan Reitti A-siipeen -korpuksen tiedostoja **reitti_a-siipeen.wav** ja **reitti_a-siipeen.mp4.** Voit ladata ääni- ja videotiedostot koneellesi Kielipankista (aineisto on nimeltään *Reitti A-siipeen -korpuksen ladattava versio* ja sen kuvailutiedot löytyvät pysyvällä tunnisteella urn:nbn:fi:lb-2020112929).

   Ko. äänitiedoston litterointia oletkin jo harjoitellut Praatilla. Litterointi on aika paljon helpompaa, kun voi katsoa videota, eikö?

   Harjoittele sen verran, että osaat lisätä ja muokata annotaatioyksiköitä (ELANissa '*annotations*') ja luoda tarpeen mukaan uusia annotaatiokerroksia (*tiers*). Harjoittele tiedoston tallentamista ja varmista, että saat sen avatuksi uudestaan ELANissa (äänineen ja videoineen päivineen). Kokeile myös viedä annotaatiotiedosto tekstimuotoon **Export As: Traditional Transcript Text…** -komennolla tai Praatin TextGrid-muotoon komennolla **Export As: Praat TextGrid…**

   Siirry sitten tehtävissä eteenpäin.

4. Lataa koneellesi äänitiedosto **pohjantuuli_M1_1.wav**. Luo ELANissa uusi annotaatiotiedosto (**File: New…**), johon lisäät mediatiedostoksi vain kyseisen äänitiedoston. Tallenna uusi annotaatiotiedosto ***.eaf***-päätteisenä.

5. Ensi alkuun riittää mainiosti yksi ainoa annotaatiokerros. Muutetaan kuitenkin *default*-annotaatiokerroksen nimi vähän paremmin kuvaavaksi. Valitse **Tier: Change Tier Attributes…**. Vaihda kerroksen nimeksi esimerkiksi *M1-clause* (tai muuta vastaavaa puhujan ja kerroksen sisällön mukaan) ja hyväksy muutos klikkaamalla alhaalta **Change**.

6. Rajaa ja litteroi äänitiedostosta tähän annotaatiokerrokseen **kaikki lauseet**. Voit tehdä rajauksen karkeasti tai tarkemmin oman mielenkiintosi mukaan. Alkuperäinen ääneen luettu tarina löytyy avuksesi tästä:

   ```
   Pohjantuuli ja aurinko

   Pohjantuuli ja aurinko väittelivät kummalla olisi enemmän voimaa, kun he
   samalla näkivät kulkijan, jolla oli yllään lämmin takki. Silloin he
   sopivat, että se on voimakkaampi, joka nopeammin saa kulkijan riisumaan
   takkinsa. Pohjantuuli alkoi puhaltaa niin että viuhui, mutta mitä
   kovemmin se puhalsi, sitä tiukemmin kääri mies takin ympärilleen, ja
   viimein tuuli luopui koko hommasta. Silloin alkoi aurinko loistaa
   lämpimästi, eikä aikaakaan niin kulkija riisui manttelinsa. Niin oli
   tuulen pakko myöntää, että aurinko oli kuin olikin heistä vahvempi.
   ```

   (Huom. Koska kyseessä on vanha kansansatu, em. tekstin käyttämiseen ei liity tekijänoikeusrajoituksia, mikä on aina hyvä asia.)

   Muista tallentaa!

7. Kun olet saanut tarinan litteroitua ainakin suurimmaksi osaksi, niin kokeillaanpa, miten ELANissa voi rakennella toisiinsa linkitettyjä annotaatiokerroksia.

   - ELAN-ohjelmassa jokainen annotaatiokerros voidaan *tyypittää*: kerrokseen voidaan liittää tieto siitä, minkätyyppistä informaatiota sen sisältämissä annotaatioyksiköissä kuvataan. Samassa tiedostossa voi olla esimerkiksi useita annotaatiokerroksia, joista jotkut sisältävät keskustelun eri osallistujien puhunnoksia, toiset taas yksittäisten sanojen kieliopillisia kuvauksia, toiset tietoa siitä, mihin kunkin osallistujan katse kohdistuu eri ajanhetkinä, jotkut taas taustamelun kuvausta, jne. Tyyppien avulla voidaan siis ryhmitellä erilaiset annotaatiokerrokset järjestelmällisellä tavalla.

   - Tyypittäminen on hakujen kannalta erityisen arvokasta. Kun kokonainen, monia annotaatiotiedostoja sisältävä korpus on annotoitu yhtenäisellä tavalla, voidaan tehdä koko korpuksesta hakuja ja kohdistaa yksittäinen haku kerralla koko korpuksen kaikkiin tietyntyyppisiin kerroksiin.

   - ELANissa voi myös käyttää tietyntyyppisissä annotaatiokerroksissa ns. kontrolloitua tai rajattua sanastoa (*controlled vocabulary*), jolloin kerroksen sisältämissä annotaatioissa on sallittua käyttää vain määrättyjä nimikkeitä. Tämä voi olla hyödyllistä tietyllä kohdealueella, esim. jos halutaan annotoida sanojen sanaluokkia, merkata äänenlaatupiirteitä tai merkitä käsien liikkeitä tietyllä systeemillä. Kun rajattu sanasto on käytössä, ELAN valvoo automaattisesti, että käyttäjä syöttää kyseiseen annotaatiokerrokseen vain sallittuja nimikkeitä.

   - ELANissa on oletuksena käytössä vain yksi annotaatiokerrosten tyyppi (*linguistic type*), joka on aluksi nimeltään tylsästi *default-lt.* Muokataan oletustyyppi kuvaavammaksi valitsemalla **Type: Change Linguistic Type…**. Tähän asti

annotoitu kerroshan sisältää lauseen tapaisia yksiköitä, joten vaihdetaan nimeksi vaikkapa *clause*. Hyväksy muutos klikkaamalla **Change**.

- Jos nyt käyt katsomassa aiemmin litteroimasi annotaatiokerroksen ominaisuuksia (**Tier: Change Tier Attributes…**), siellä pitäisi lukea Linguistic Type -kohdassa "clause", jos tyypin muuttaminen onnistui.

8. Seuraavaksi yritetään luoda uusi tyyppi sellaisia annotaatiokerroksia varten, joihin halutaan erikseen annotoida litteroitujen lauseiden sisältämiä sanoja.

    - Valitse **Type: Add New Linguistic Type…**

    - Kirjoita Type Name -kohtaan esimerkiksi *word.*

    - Haluamme myös, että word-tyyppiset annotaatioyksiköt esiintyvät aina *clause*-tyyppisten yksiköiden rajojen sisäpuolella. Tällaisen annotaatiokerrosten välisen suhteen voit määritellä valitsemalla *Stereotype*-kohdasta vaihtoehdon *Included In.*

    - Klikkaa lopuksi **Add.**

      Lisävinkki:
      Jos olet aikeissa lisäksi annotoida vaikkapa jokaisen sanan sanaluokan omaan annotaatiokerrokseensa, kannattaa sanaluokille luoda oma tyyppi. Se tehdään muuten samoin kuin edellä, mutta Stereotype-kohdasta pitää valita *Symbolic Association* (mikä tarkoittaa, että jokaista "emokerroksen" annotaatiota vastaa yksi "tytärkerroksen" annotaatio). Tyypin nimeksi kannattaa toki myös kirjoittaa esim. *part of speech* tai *POS* eli 'sanaluokka' tms.

      Jos/kun sanaluokan merkitsemiseen käytetään yleensä tarkoin rajattua määrää nimikkeitä (esim. verbi, substantiivi, adjektiivi, numeraali jne.), näille voidaan jo etukäteen määritellä rajattu sanasto (*controlled vocabulary*). Halutessasi voit kokeilla sellaisen tekemistä **Edit: Edit Controlled Vocabularies…** Anna sanastolle aluksi nimi (*CV Name*) ja klikkaa *Add*. Sanaston sisältämät nimikkeet ja niiden kuvaukset lisätään tämän jälkeen ikkunan alareunassa yksi kerrallaan. Luotu sanasto pitää sitten myös ottaa käyttöön sanaluokka-tyypissä (**Change Linguistic Type:** Use Controlled Vocabulary). Tämän jälkeen ELAN auttaa sinua valitsemaan sanaluokaksi valmiilta listalta aina jonkin itse määrittelemistäsi nimikkeistä.

9. Nyt koetetaan automaattisesti *saneistaa* eli jakaa yksittäisiin sanoihin litteroimasi *M1-clause*-niminen annotaatiokerros niin, että luodaan samalla uusi annotaatiokerros, jonka tyypiksi annetaan äsken määritelty *word:*

    - Valitse **Tier: Tokenize Tier…** ("tokenisointi" tarkoittaa tekstin jakamista saneisiin).
    - Kohdassa *Source tier (parent tier)* pitäisi olla valittuna alkuperäinen, *M1-clause*-niminen kerros.
    - Klikkaa kohdan *Destination tier* oikealla puolella olevaa painiketta Create New Tier…
    - Luodaan uusi kerros, johon yksittäiset saneet tulevat. Anna kerroksen nimeksi vaikkapa *M1-word.* Valitse *Parent Tier* -kohdasta alkuperäinen *M1-clause*-kerros.

Valitse *Linguistic Type*-kohdasta *word*. Hyväksy lopuksi klikkaamalla **Add** ja sulje ikkuna klikkaamalla **Close**. Uusi M1-word-niminen kerros ilmestyy taustalle, mutta se on vielä tyhjä.
    - Jatka saneistamista klikkaamalla **Tokenize Tier**-ikkunassa Start.
    - Uuteen annotaatiokerrokseen ilmestyy nyt automaattisesti alkuperäisessä litteraatissa esiintyviä sanoja vastaavat sanayksiköt. Kätevää! (Sanarajat tulevat toki automaattisesti tasavälein, eli ne eivät kohdistu oikeisiin kohtiin äänitiedostossa, joten mikäli tarkkoja sanarajoja tarvitaan tutkimuksessa, rajaukset joutuu käsin tarkistamaan ja siirtämään kohdilleen.)
    - Tässä välissä kannattaa taas tallentaa tiedosto!

10. Jos haluat annotoida jokaisen sanan sanaluokan omaan kerrokseensa ja olet jo luonut lingvistisen tyypin sanaluokkaa varten (ks. kohta 8 edellä), voit nyt luoda uuden kerroksen puhujan M1 lukemien sanojen sanaluokitukselle:
    - Valitse **Tier: Add New Tier…**
    - Anna kerroksen nimeksi vaikkapa *M1-POS*.
    - Valitse uuden sanaluokkakerroksen "emokerrokseksi" äsken putkautettu sanakerros, *Parent Tier: M1-word*.
    - Valitse *Linguistic Type: part-of-speech* (tai *POS*).
    - Hyväksy klikkaamalla **Add.**
    - Sanaluokkakerroksiin luotavat yksiköt seuraavat nyt aina niitä vastaavien sanayksiköiden rajoja, ts. niitä ei voi siirrellä itsenäisesti.

11. Mieti, millaisia annotaatiokerroksia itse tarvitsisit ja millaisia suhteita niiden välille voisi annotaatiokerroksia tyypittämällä määritellä ELANissa.

12. Entäpä jos haluaisit tehdä hakuja annotoimastasi aineistosta? ELANin **Search**-valikosta löytyy useita hakutoimintoja:
    - **Find (and Replace)…,** haku avoinna olevan tiedoston sisältä
    - **Search Multiple eaf…**, yksinkertainen merkkijonohaku omalla koneellasi olevista eaf-tiedostoista tai hakemistoista
    - **Structured Search Multiple eaf…**, monikerroksinen haku omalla koneellasi olevista eaf-tiedostoista tai hakemistoista

    Huom. Multiple eaf -toimintojen aluksi sinun on määritettävä **hakualue**, *domain*, jossa oleviin EAF-muotoisiin annotaatiotiedostoihin haut kohdistuvat. Hakualueeseen voit poimia haluamasi joukon omalla koneellasi olevia yksittäisiä eaf-tiedostoja ja/tai kokonaisia hakemistoja. Hakualueen voi tallentaa sopivalla nimellä ja käyttää sitä myöhemmin uudestaan, ts. näin voit määritellä käyttöösi kokonaisia puhekorpuksia.

13. Kokeile tehdä ELANilla erilaisia hakuja itse litteroimastasi *Pohjantuuli ja aurinko* -tarinasta (tai muista ELANilla annotoimistasi tiedostoista). ELAN-ohjelman manuaalista löytyy tarkkoja ohjeita hakujen tekemiseen.

    Kun määrittelet haun kohteena olevaa sanaa tms. annotoitua merkkijonoa, voit hyödyntää myös ns. **säännöllisiä lausekkeita** (*regular expressions*). Niiden avulla voi määritellä esimerkiksi haettavan annotaation alku- tai loppuosan, annotaation sisällä esiintyviä tai toistuvia merkkejä, merkkejä, jotka eivät saa esiintyä haettavissa kohteissa jne. Luettelo kaikista ELANin tukemien säännöllisten lausekkeiden operaattoreista löytyy

täältä: http://www.mpi.nl/corpus/html/elan/apa.html

Voit myös kysyä lisää kurssialueen keskustelufoorumilla, niin pohditaan yhdessä!

Hakuehdotuksia:

- Etsi sanoja, joiden litteroitu muoto sisältää pitkän vokaalin *ää*.
- Etsi sanoja, jotka esiintyvät lauseen (tai muun rajaamasi yksikön) alussa (tai lopussa).
  - Vinkki: annotoidun merkkijonon alkua voi merkitä säännöllisissä lausekkeissa merkillä ^ ja loppua merkillä **$**. Sananrajaa merkitään **\b**.
  - Mikä tahansa sana voitaisiin mahdollisesti ilmaista säännöllisenä lausekkeena vaikkapa näin:
    \b[a-zääöA-ZÅÄÖ]*\b
    = sanarajojen välissä esiintyvä, kirjaimista a-ö tai A-Ö koostuva mikä tahansa merkkijono. (Tähti tarkoittaa, että hakasulkeissa mainittuja merkkejä voi olla miten monta tahansa.)
- Jos olet annotoinut sanaluokkia, kokeile etsiä vaikkapa kaikki token-tyyppisten (*Tier type: token*) annotaatiokerrosten sisältämät sanayksiköt, joiden kohdalla sanaluokkakerroksessa lukee **V** (verbi) ja joiden alkukirjain on **o**
  - Huom. tämä on jo aika hankala haku! Tulokseksi pitäisi tulla lähinnä olla-verbin eri muotoja.

14. Nyt olet saanut käsityksen siitä, mitä ELANilla voi tehdä ja kuinka "tavallista" litterointia voi helposti **rikastaa**.
    - Pohdi lopuksi, mihin tarkoitukseen ELANista on eniten hyötyä ja missä asioissa Praat on parempi.
    - Kumpaa uskot tarvitsevasi jatkossa enemmän, Praatia vai ELANia?
    - Voitko käyttää molempia ohjelmia rinnakkain? Jos haluat, voit kokeilla, kuinka onnistuu Praatin TextGrid-muotoisen annotaatiotiedoston tuonti ELANiin, **Import: Praat TextGrid File…**

Onnittelut ahkeroinnista – toivottavasti siitä on ollut sinulle hyötyä!