Pirkko Suihkonen

UHLCS:n korpusten siirtäminen CSC:n corpus-koneelle: yhteenveto työvaiheista (10.9.2007)

1. Työn alkuvaihe

Työn ensimmäisessä vaiheessa selvitettiin, mitä korpuksia on, millaisia sopimuksia korpuksista on, ja miten korpukset on dokumentoitu. Kirjoitin työn edistymisestä väliaikaraportteja, jotka liitin projektin KitWiki-tiedostoon. Raportit ovat olleet pdf-muodossa. Pdf-muoto oli kätevä siksi, että useimmat tiedostot sisältävät taulukoita, joita muokattiin sitä mukaa, kun työ edistyi. Raporttien päivitetyt versiot ovat projektin KitWiki-sivuilla.

2. Korpusten sopimukset ja yhteydenpito korpusten omistajiin

Työn koko keston aikana oltiin yhteydessä korpusten omistajiin tekijänoikeuksia koskevien sopimusten päivittämiseksi ja joissakin tapauksissa myös niiden selvittämiseksi. Korpusten omistajia informoitiin korpusten siirrosta, ja omistajiin luvattiin ottaa yhteyttä, kun korpukset on siirretty. Kaikkia korpusten omistajia on informoitu siitä, että korpukset ovat nyt käytettävissä. Hankalia käsitellä olivat esim. venäjän korpukset, joista ainoastaan Tampereen korpuksesta oli saatavilla yksityiskohtainen sopimus. Muiden korpusten joukossa oli useita pieniä korpuksia, joista ainakaan toistaiseksi ei ole sopimuksia (esim. jiddish ja ketshua). Mahdollisuutta saada aikaan ketshuan korpusta koskeva sopimus selvitetään vielä, mutta jiddishin korpus jätetään laitoksen tutkijoiden käyttöön korpuksen omistajan suullisen luvan perusteella. Sama koskee latinan korpusta. Useista suurista englannin korpuksista ei löytynyt voimassa olevaa sopimusta. Korpusten sopimuksia koskevia kysymyksiä kokosin tiedostoon ”corp-contracts.pdf”, joka tarkentui sitä mukaa kun työssä edettiin. Tiedoston eri versioita oli käytettävissä työn aikana.

3. Korpusten metadata-tiedostot ja UHLCS:n info-sivut

Seuraavassa vaiheessa korjasin ja päivitin korpusten metadata-tiedostot. Tiedostoihin on laitettu tieto korpusten nykyisestä sijainnista. Metadata-tiedostojen yksityiskohtainen tarkastaminen ja korjaus oli työlästä, koska tiedostoissa olevia virheitä ei aina ollut helppo havaita. Metadata-tiedostot sijaitsevat UHLCS:n info-sivujen yhteydessä osoitteessa http://www.ling.helsinki.fi/uhlcs/metadata/corpus-metadata/. Tiedostojen kopiot sijaitsevat korpushakemistoissa korpusten yhteydessä. Korjasin, oikoluin ja päivitin myös UHLCS:n info-sivut, laadin info-sivujen suomenkielisen version, sekä päivitettiin yhteyshenkilöiden osoitetiedot. Korpusten info-sivut ovat nähtävillä osoitteessa http://www.ling.helsinki.fi.

4. Korpusten käyttäjien ryhmät angarakilla

Korpusten statuksen selvittämiseen tarvittiin myös tietoa niistä ryhmistä, jotka saivat korpuksia käyttää. Selvitin eri korpushakemistoissa olevat käyttäjien ryhmät siinä määrin kuin se oli mahdollista. Kävi ilmi, että vaikka korpuksen tai alakorpuksen hakemisto oli avoinna sille ryhmälle, johon kuuluin, oli hakemistojen sisällä paljon sellaisia hakemistoja, jotka olivat suljettuja. Korpusten ryhmien kohdalla oli myös usein numerosarjoja. Näistä monet edustivat sellaisia ryhmiä, joita ei enää ollut olemassa, mistä syystä hakemistot olivat suljettuja. Myös korpusten omistajien kohdalla esiintyi numerosarjoja, ja sen selvittämiseen, mitä näiden numerosarjojen takana oli, tarvittiin hallinnon apua (admn@ling.helsinki.fi). Laadin hakemistoista ja ryhmistä taulukkomuotoisen tiedoston, johon kokosin lisäksi tietoa ryhmien jäsenistä käyttämällä apuna hakemistossa /etc olevaa ryhmiä koskevaa tiedostoa. Tietoa korpusten käyttäjien ryhmistä angarakilla hyödynnetään siinä vaiheessa, kun määritellään ryhmät niille käyttäjille, joilla on jo käyttölupa CSC:llä.

5. Korpusten käyttäjien ryhmät CSC:n corpus-koneella

Kimmo Koskenniemen kanssa käydyn keskustelun perusteella laadin uuden ryhmäjaon, joka ensimmäisessä vaiheessa perustui korpusten auktorisointiin, sitten korpusten käyttäjäryhmiin, ja kolmanneksi työn tarkoitukseen. Ryhmät muodostettiin sen perusteella, miten nämä kolme osatekijää saattoivat kombinoitua. Alkuperäisessä versiossa kukin osatekijä erotettiin tavuviivalla, lopullisessa versiossa CSC:llä oli poistettu käyttäjäryhmän ja käyttötarkoituksen välinen tavuviiva. Ryhmäjaosta tuli kompleksinen, ja Eero Vitien kanssa sovittiin, että CSC:llä muodostetaan ryhmiä sitä mukaa kuin niistä esiintyy aktuaalista tarvetta. Joitakin kommentteja ryhmäjakoon on muistiossa 1.9.2007, joka on KitWikin verkkosivulla.

6. Korpusten siirrosta

Korpukset siirrettiin useassa vaiheessa. Ensimmäisen kerran siirrettiin ne korpukset, jotka ovat minun vastuullani, jo ensimmäisen tapaamisen yhteydessä CSC:llä. Tässä yhteydessä siirretyissä korpuksissa oli kuitenkin siinä määrin työstettävää, että katsoin parhaaksi tehdä koko työn angorakin kautta. Eri vaiheiden jälkeen siirrettiin uudelleen muokatut korpukset Jukka Huhdan avustuksella 28.8.20007. Siirtämättä jäi muutamia hallinnon omistamia hakemistoja sekä hakemistossa /corp/fin olevia hakemistoja, joiden omistajia halusin informoida korpusten siirrosta. Näistä aineistoista on informoitu Eero Vitisetä ja osoitetta admin@ling.helsinki.fi. Mm. Sari Salmisuon ja Kimmo Koskenniemen aineistoja on vielä siirtämättä (ks. erillinen luettelo näistä hakemistoista: ”hakemistot-joita-ei-ole-siirretty-8-9-2007.pdf”).

4.7.2007 CSC:llä pidetyssä kokouksessa sovittiin, että CSC:lle siirretään vain ne korpukset, joiden sopimukset ovat kunnossa. Töiden käytännön järjestämisen kannalta oli

selkeämpää, että siirrettiin kerralla jokseenkin kaikki korpukset. Tämä mahdollisti myös sen, että kaikkia korpuksia voitiin käsitellä yhtenäisesti. Kimmon kanssa sovittiin suullisesti, että siirretään kaikki korpukset, mutta korpukset laitetaan vasta käyttöön, kun niiden sopimukset sen sallivat. Nyt ne korpukset, joiden sopimusten selvitys on vielä kesken, ovat CSC:n corpus-koneella hakemistossa ADM. Jos korpusten sopimukset eivät selviä, korpukset poistetaan corpus-koneelta.

7. Korpusten siirron aikana pidetyt suunnittelukokoukset

Yleisen kielitieteen laitoksella 18.6.207 ja CSC:Llä 4.7.2007 pidettyjen tapaamisten lisäksi tapasimme Eero Vitien kanssa CSC:llä kolmesti (24.8, 27.8. ja 5.9.2007). Tapaamisissa keskusteltiin korpusten siirtoon ja käyttöönottoon liittyvistä yksityiskohdista. Tapaamisista on muistiot KitWikin verkkosivuilla. Tapaamisissa käytyjä keskusteluja ja niistä tehtyjä muistioita on pidetty ohjeellisena korpusten siirtoa ja käyttöönottoa koskevissa kysymyksissä, ja Eero Vitien kanssa käydyistä keskusteluista tehdyt raportit ovat osa työnkulusta tehtyä yhteenvetoa.

8. Työajasta

Jos korpusten siirto olisi koskenut ainoastaan korpuksia, joiden sopimuksista minä olin vastuussa, työaika olisi ollut liian pitkä. Vaikka korpusten omistajien kanssa oli tarpeen käydä yksityiskohtaista keskustelua korpusten siirrosta ja työn jatkamisesta, ja joissakin tapauksissa keskustelun kulku oli hidasta, olivat kaikki peruslähtökohdat kuitenkin selvät. Sen sijaan kaikkien korpusten siirtämiseksi työaika loppui kesken. Nyt tilanne on kuitenkin siinä määrin selvä, että tästä on mahdollista jatkaa eteenpäin.

9. Keskeneräinen osuus (10.9.2007)

ADM-hakemistossa olevien aineistojen ja siirtämättömien listassa olevien aineistojen auktorisoinnin selvittäminen vaatii lisäaikaa ja lisätyötä. Anssi huolehtii hakemistossa /ADM/bible olevien korpusten sopimuksista, ja hakemistossa /ADM/russian olevista aineistoista on tiedusteltu korpusten omistajilta tai toimittajilta. Huolehdin siitä, että sitä mukaa kuin saan tietoa hakemistossa /ADM/russian olevien korpusten sopimuksista, tämä tieto tulee myös otetuksi huomioon näiden korpusten käyttöönottoon liittyvien kysymysten päivittämisessä. Hakemistoissa /ADM/english, /ADM/finnish, /ADM/german, /ADM/swedish, /ADM/somali ja /ADM/spoken olevien aineistojen auktorisointi on parhaiten selvitettävissä laitoksen kautta.